Environmental Modelling and Software 110 (2018) 72–94

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A knowledge modelling framework for intelligent environmental decision support systems and its application to some environmental problems

Mihaela Oprea

Petroleum-Gas University of Ploiesti, Department of Automatic Control, Computers and Electronics, Bdul Bucuresti No 39, Ploiesti, Romania

ARTICLE INFO

Keywords:
Knowledge modelling framework
Integrated environmental modelling
Bayesian networks
Data mining
Ontology
Intelligent environmental decision support system

ABSTRACT

Environmental processes are highly complex and their understanding involves the analysis of various quantitative and qualitative parameters (physical, chemical, geographical etc), which are more or less correlated. Appropriate environmental knowledge can deal with this complexity in a tractable way. Such knowledge is essential for solving particular environmental problems. Generating valuable environmental knowledge is a challenging research topic, especially for environmental data science, as efficient knowledge can lie behind data. Integrated environmental modelling uses a holistic view and can provide a possible better solution to environmental problems understanding. The paper presents a knowledge modelling framework for intelligent environmental decision support systems (IEDSS) by following such a holistic perspective. Thus, the proposed framework integrates an ontological approach and two data analysis approaches (data mining and Bayesian networks), which are applied for the generation of a knowledge base that is used by an IEDSS for decision making. The application of the framework is illustrated on three case studies from different environmental domains: (1) water (river resource management, river water pollution analysis), (2) air (air pollution analysis, ozone prediction), and (3) soil (soil pollution analysis).

1. Introduction

One of the current challenging research directions in environmental sciences is integrated environmental modelling (IEM) (see e.g. (Laniak et al., 2013)), that can be viewed as a possible better solution to the environmental processes understanding problem. IEM applies a holistic view on environmental problems solving. In this context, various data analysis methods and techniques (as e.g. statistical and artificial intelligence-based), can be combined and applied in order to provide valuable environmental knowledge for the decision making process of an intelligent environmental decision support system.

Several artificial intelligence (AI) methods can be used for environmental systems modelling (see e.g. the overviews presented in (Chen et al., 2008) and (Struss, 2008)). Examples of such AI-based methods are knowledge-based methods (e.g. knowledge based systems, expert systems, case based reasoning) and computational intelligence methods (e.g. fuzzy inference system, artificial neural networks, genetic algorithms, and swarm intelligence methods). These methods can be used for integrated environmental modelling in decision support systems. Many environmental problems can be tackled with intelligent tools such as IEDSS, a variety of IEDSS applications being reported in the literature, as for example, MODULUS DSS - for land degradation in the Mediterranean (Oxley et al., 2004), an IEDSS based on multi-agent systems for water management (Urbani and Delhom, 2005), i-EKbase - an IEDSS for sustainable agriculture (Dutta et al., 2014), FLIRE DSS - a web-based DSS for floods and wildfires in urban and periurban areas (Kochilakis et al., 2016). An overview on key challenges and best practices in environmental decision support systems development is presented in (McIntosh et al., 2011), while selected worldwide developed IEDSS tools are briefly described in (Gibert et al., 2012). Some frameworks for IEDSS development were also reported in the literature, an example being given in (Sànchez-Marrè et al., 2008a).

Other promising model-based approaches for EDSS development include qualitative models such as the qualitative reasoning model for algal bloom in the Danube Delta Biosphere Reserve proposed in (Cioaca et al., 2009), the qualitative dynamic model adapted for hydroecology, which is presented in (Heller and Struss, 2001), the abductive reasoning models described in (Wotawa et al., 2010) and (Wotawa, 2011), and other AI-based models (see e.g. the models discussed in (Sànchez-Marrè et al., 2008b) and (Struss et al., 2003)).

The purpose of our research work is to provide a knowledge modelling framework for intelligent environmental decision support systems. The paper presents an environmental knowledge modelling framework that integrates an ontological approach and two data analysis approaches (data mining and Bayesian networks), which are applied for the generation of a knowledge base used by an IEDSS for decision

E-mail address: mihaela@upg-ploiesti.ro.

https://doi.org/10.1016/j.envsoft.2018.09.001
Received 24 February 2017; Received in revised form 19 July 2018; Accepted 5 September 2018
Available online 15 September 2018
1364-8152/

making. Starting from a conceptual model of the environmental problem it is developed the problem domain ontology used in the next steps when data mining and Bayesian networks are applied to generate rules from data sets and decision tables. The application of the framework is shown on three case studies for solving different environmental problems in the domains of water, air and soil.

2. Literature review

A brief literature review on the ontological approach, data mining and Bayesian networks application in environmental science is presented.

2.1. Ontological approach in environmental science

The ontological approach tackles expertise domain knowledge conceptualization, providing a domain ontology, which includes a vocabulary with terms (concepts from general to particular ones, and relations between concepts), and axioms describing restrictions on concepts and relations. Actually, it performs a domain knowledge modelling. Various applications of ontological approach in environmental applications are given in the literature (e.g. wastewater management with ontology-based EDSS, OntoWEDSS (Ceccaroni et al., 2004), knowledge modelling in an air pollution control decision support system (Oprea, 2005), flow and water quality modelling by using an ontology-based knowledge management system (Chau, 2007), integration of air quality models and 3D city models with ontologies (Metral et al., 2008), air pollution ontology (Czarnecki and Orłowski, 2009), simulation in agricultural systems modelling based on ontology (Beck et al., 2010), environmental impact assessment (EIA) ontology covering the environmental experts terminology of various aspects such as water, air, soil, habitat, geophysical, and socioeconomic impact (Garrido and Requena, 2011), description and classification of USDA soil taxonomy up to soil series (Deb et al., 2015), description of learning resources on organic agriculture with an AGROVOC-based ontology - soil pollution analysis due to fertilization (Sánchez-Alonso and Sicilia, 2009), soil properties and processes ontology, OSP (Du et al., 2014), knowledge modelling in river water quality monitoring and assessment (Xiaomin et al., 2016)).

2.2. Data mining in environmental science

Data mining (DM) integrates techniques from several domains (as e.g. statistics, machine learning, and artificial intelligence), and performs knowledge extraction from large data bases. Examples of DM techniques (Cios et al., 2007) are decision trees, rule algorithms, hybrid algorithms, artificial neural networks, and statistical methods (e.g. regression). A brief description and classification of DM techniques oriented to decision making and some guidelines regarding the selection of the right technique are described in (Gibert et al., 2010a). Some models and algorithms for decision making in a data driven environment are discussed in (Kusiak, 2002). The DM approach was used in several environmental applications, as e.g. water supply assets modelling (Babovic et al., 2002), air pollution management policy making (Li and Shue, 2004), simulating farmers' crop choices for integrated water resource management (Ekasingh et al., 2005), water quality management using GIS data mining (Karimipour et al., 2005), analysis of water pollution effects on human health - the results are used for water management by governmental authorities (Sokolova and Fernández-Caballero, 2007), agricultural soil profiles characterization (Armstrong et al., 2007), air pollution monitoring and mining based on sensor grid (Ma et al., 2008), air pollution modelling and short-term air quality forecasting (Riga et al., 2009), contaminant event locations identification in water distribution systems (Huang and McBean, 2009), soil carbon mapping in Australia (Bui et al., 2009), knowledge modelling in a waste water treatment plant by explicit knowledge discovery from the corresponding dynamic processes with the clustering based rules by states approach (Gibert et al., 2010b), water quality monitoring using remote sensing data mining (Wen and Yang, 2011), analysis of air pollution effects on humans, including geographic heterogeneity assessment (Stanley Young and Xia, 2013), air quality monitoring (Czechowski et al., 2013), ground water quality assessment (Kolli and Seshadri, 2013), soil organic matter prediction (Teixeira et al., 2014), soil property variation mapping by applying data mining to soil category maps (Du et al., 2014), air pollution prediction (Siwek and Ossowski, 2016).

2.3. Bayesian networks in environmental science

Bayesian networks are directed acyclic graphs that represent the dependences between a set of variables and their joint probability distribution, being probabilistic graphical models, which integrate quantitative and qualitative data. They are used in environmental modelling proving to be a useful support instrument for decision making in various environmental problems (see e.g. urban air pollution forecasting (Cossentino et al., 2001), farm irrigation in (Wang et al., 2009) and (Robertson et al., 2009), simulation of water flow in organic soils, based on ontology (Kwon et al., 2010), groundwater management for agriculture (Carmona et al., 2011), air pollution related health risk assessment (Liu et al., 2012), GHG gas emissions management in the British agricultural sector (Pérez-Miñana et al., 2012), soil pollution assessment (Li et al., 2015), water resource management in (Phan et al., 2016), ecological risk assessment in estuarine ecosystems in (McDonald et al., 2016), emergent water pollution accidents risk analysis in (Tang et al., 2016), air pollution prediction via multi-label classification (Corani and Scanagatta, 2016)). An overview on Bayesian networks use in environmental modelling is presented in (Aquillera et al., 2011), another overview on BNs use in environmental and resource management is described in (Barton et al., 2012), while some good practice guidelines are synthesized in (Chen and Pollino, 2012). Also, it is important to note that Bayesian networks can be generated by using the domain ontology (see e.g. Fenz et al., 2009).

2.4. Review synthesis

Table 1 shows a synthesis of the literature review made for different environmental problems solving with the three approaches: data mining, Bayesian networks and ontological approach, within the framework of environmental decision support systems or knowledge modelling.

The main conclusion of the literature review is that all three approaches were applied with success to different environmental problems solving (pollution, prediction, resource management, monitoring, control, impact assessment - water, air and soil), and have an important potential in generating valuable environmental knowledge. However, none of the studied problems were solved by applying all three approaches.

3. Proposed knowledge modelling framework

3.1. Our approach

Starting from the literature review, we have selected some artificial intelligence based methods (data mining, knowledge based systems) that proved to tackle in a proper manner the problem of knowledge modelling in an IEDSS. As environmental processes are very complex

Table 1
Synthesis of literature review (selected references).

Method	Environmental problem		
	Water	Air	Soil
Data mining	Analysis of water pollution effects on human health - results are used for water management (Sokolova and Fernández-Caballero, 2007); Water quality monitoring using remote sensing data mining (Wen and Yang, 2011); Water quality management using GIS data mining (Karimipour et al., 2005); Ground water quality assessment (Kolli and Seshadri, 2013); Identify contaminant event locations in water distribution systems (Huang and McBean, 2009); Waste water treatment plant knowledge modelling (Gibert et al., 2010b); Farmers' crop choices simulation for integrated water resource management (Ekasingh et al., 2005); Water supply assets modelling (Babovic et al., 2002);	Air pollution modelling and air quality prediction (Riga et al., 2009); Air pollution prediction (Siwek and Ossowski, 2016); Analysis of air pollution effects on humans (Stanley Young and Xia, 2013); Air pollution monitoring and mining based on sensor grid (Ma et al., 2008); Air quality monitoring (Czechowski et al., 2013); Policy making in air pollution management (Li and Shue, 2004);	Soil organic matter prediction (Teixeira et al., 2014); Soil property variation mapping by applying data mining to soil category maps (Du et al., 2014); Soil carbon mapping (Bui et al., 2009); Agricultural soil profiles characterization (Armstrong et al., 2007);
Bayesian networks	Water resource management (Phan et al., 2016); Groundwater management - water and soil (Carmona et al., 2011); Ecological risk assessment in estuarine ecosystems (McDonald et al., 2016); Emergent water pollution risk analysis (Tang et al., 2016);	Urban air pollution forecasting (Cossentino et al., 2001); Air pollution prediction (Corani and Scanagatta, 2016); Air pollution related health risk assessment (Liu et al., 2012);	Soil pollution assessment (Li et al., 2015) Farm irrigation - soil and water (Wang et al., 2009), and (Robertson et al., 2009); GHG gas emissions management in the British agricultural sector - farming (Pérez-Miñana et al., 2012);
Ontology	River water quality monitoring and assessment (Xiaomin et al., 2016); Ontology-based simulation of water flow in organic soil (Kwon et al., 2010); Flow and water quality modelling (Chau, 2007); Wastewater management (Ceccaroni et al., 2004); EIA ontology - assessment of water impact (Garrido and Requena, 2011);	Air quality models and 3D city models integration with ontologies (Metral et al., 2008); Air pollution ontology (Czarnecki and Orłowski, 2009); Air pollution control (Oprea, 2005); EIA ontology - atmosphere impact assessment (Garrido and Requena, 2011);	Soil properties and processes ontology, OSP (Du et al., 2014); Soil taxonomy up to soil series description and classification (Deb et al., 2015); Simulation in agricultural systems modelling based on ontology (Beck et al., 2010); Organic agriculture learning resources description with an AGROVOC-based ontology (Sánchez-Alonso and Sicilia, 2009); EIA ontology - ground and landscape impact assessment (Garrido and Requena, 2011);

with several uncertainties and various variables, we have chosen Bayesian networks to manage knowledge uncertainty in such cases. Moreover, Bayesian networks handle heterogeneous data: quantitative and qualitative, being very useful in environmental problem description.

The paper proposes an environmental knowledge modelling framework, which integrates three approaches for knowledge modelling: ontological approach, data mining and Bayesian networks, to be used in an IEDSS, as shown in Fig. 1. The ontological approach allows problem domain conceptualization and provides concepts and relations that will be used in knowledge representation and when applying the other two approaches, data mining and Bayesian networks. The data mining approach (in our case, rule induction – inductive learning techniques: decision trees and rules algorithms) performs problem specific knowledge discovery from the available data sets (i.e. problem specific data bases) and decision tables. Bayesian networks are used to represent the probable dependences between different parameters (as e.g. cause-effect relations) that describe the problem.

In our framework, knowledge is represented under the rule form, i.e. IF < premise > THEN < conclusion > , and knowledge uncertainty modelling is achieved with a probabilistic model (Bayesian networks), a confidence factors model (that quantifies the uncertainty with numerical values associated to rules generated, for example, with inductive learning DM techniques) and a fuzzy model (linguistic values associated to facts included in rules, either in the premise or in the conclusion - corresponding to each analyzed parameter specific ranges). Temporary databases (T_DB), associated to the current environmental problem context, are added to the IEDSS databases (DB).

The knowledge bases built by our approach are KB_{Domain} (include domain knowledge under the form of rules and decision tables), KB_{DM} (rules derived via inductive learning from available databases - T_DB, and decision tables), KB_{BN} (probabilistic rules derived from Bayesian networks), and the final knowledge base, $KB_{Problem}$ in which are integrated all rules necessary for the decision making procedure of the IEDSS. The knowledge included in KB_{Domain} and KB_{DM} are using fuzzy model and confidence factors model for knowledge uncertainty representation. The knowledge included in KB_{BN} is represented as probabilistic rules. Finally, the problem solving knowledge base, $KB_{Problem}$ is using all the three models, fuzzy model, confidence factors and probabilistic model, for the knowledge uncertainty representation. The main problem with the mix of these three models regards rules uniformity in the case of probabilistic rules integration with the rules using the confidence factors model, which needs further investigation. Our temporary solution was to keep all rules (probabilistic and with confidence factors) with human expert approval and possible correction, and to apply all of them during reasoning, selecting the most probable conclusion and the conclusion with the highest confidence factor, finally, the human expert deciding which one will be used by the decision making procedure depending on a specific problem's scenario context.

The basic components of a decision support system are: data analysis and decision making. The first component, *data analysis*, provides a part of the environmental knowledge (mainly from collected data

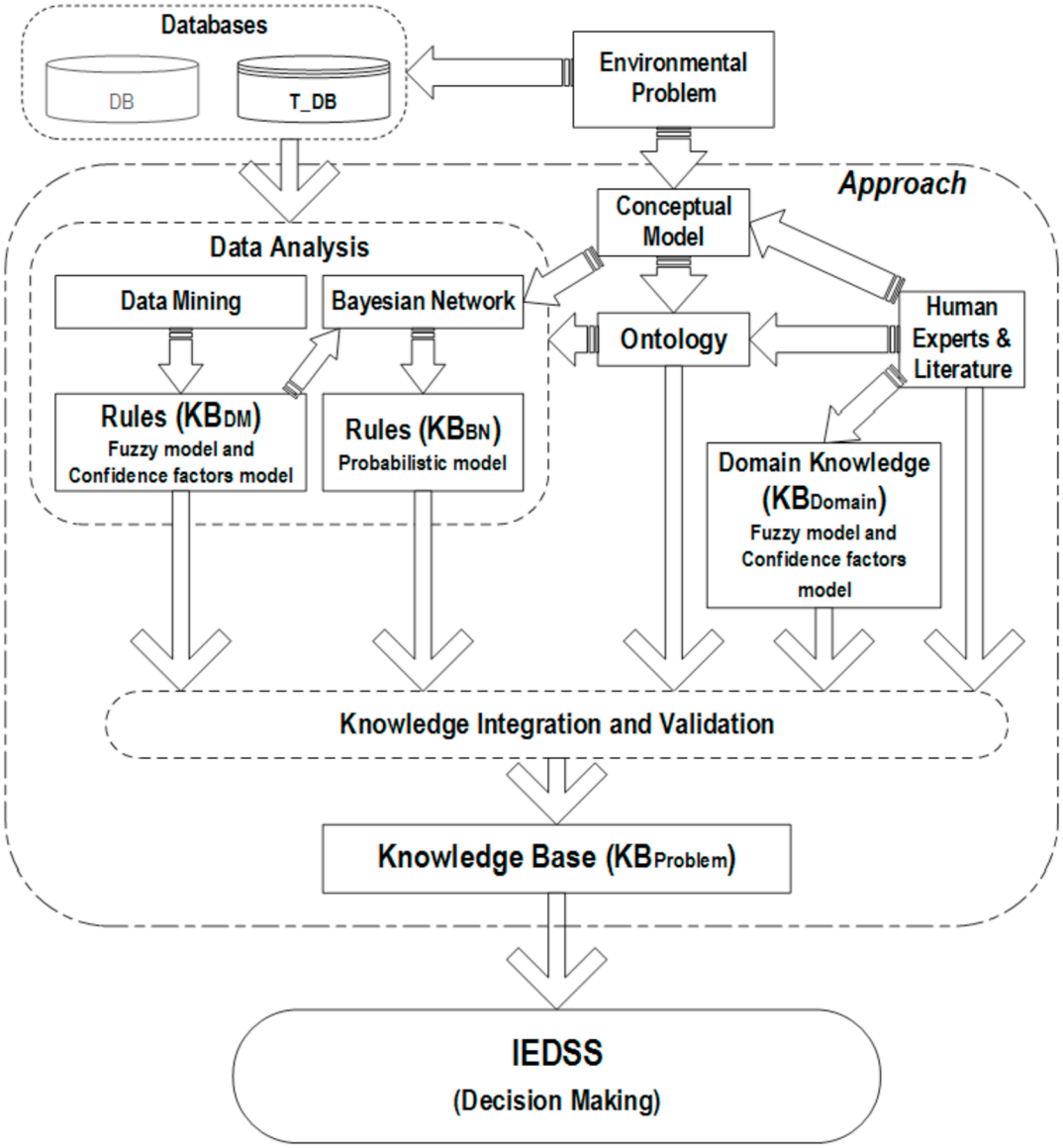

Fig. 1. Environmental knowledge modelling approach overview.

sets), while the second one, *decision making* uses this knowledge and additional problem domain knowledge for decision making.

The framework input is given by the environmental data sets (collected data, T_DB and DB) and environmental problem domain specification, while the output is given by the knowledge base, which contains the generated rules. The quality of the environmental data used by the framework will strongly influence the derived knowledge base quality and thus, the quality of the decision made by IEDSS. Therefore, data cleaning, checking and pre-processing are required to be performed within data analysis. A set of valuable guidelines for the application of pre-processing techniques in the context of environmental data mining is presented in (Gibert et al., 2016).

The framework can be applied to different environmental problems (e.g. environmental resource management, monitoring, analysis and control of environment quality - i.e. quality of water/air/soil).

3.2. Framework description

The proposed knowledge modelling framework has seven steps and starts with the problem *analysis phase* (step 1), in which the problem solving domain is identified and a conceptual model of the problem is built, followed by *the phase of problem domain knowledge conceptualization* (i.e. problem domain ontology building, step 2), and *the phase of knowledge base construction* via three approaches (steps 3, 4, 5, 6), knowledge acquisition from literature and human experts, via *inductive learning* data mining (*decision trees and rule algorithms*), and through Bayesian networks, the results of the last methods being validated by human experts and using the problem domain ontology, through reasoning. The last phase of the framework is *knowledge integration in the IEDSS knowledge base and validation* (step 7) through reasoning by using domain ontology and human experts.

KM-Framework(*Input*: Problem, Domain, DataSet; *Output*: $KB_{Problem}$)

1. $CM_{Problem}$ = *ConceptualModelDesign*(Problem);

2. $Onto_{Domain}$ = *OntologyDevelopment*(Problem, Domain, $CM_{Problem}$);

3. KB_{Domain} = *KnowledgeAcquisition*(Domain, $Onto_{Domain}$); // manually – general KB

// apply **a proper inductive learning data mining algorithm** to discover knowledge specific

// to an area/site - rules and correlations

4. KB_{DM} = *KnowledgeDiscovery*(DataSet, DMAlgorithm, $Onto_{Domain}$); // area-specific KB

5. $BN_{Problem}$ =*BayesianNetworkDevelopment*(Problem, $CM_{Problem}$, KB_{DM}, KB_{Domain}, $Onto_{Domain}$);

// * identify the most representative *n* scenarios of the problem;

5.1 *ProblemScenariosIdentification*(Problem; S, n)

// * develop a set of Bayesian networks, one for each scenario

5.2 $BN_{Problem}$ = ∅; // initialize with the empty set

5.2 for *each scenario, S_i, *i=1*, …, *n* **do**

5.2.1 CM_{Si} = *ConceptualModelDesign*(S_i);

5.2.2 BN_{Si} = *BN_Development*(CM_{Si}, KB_{DM}, KB_{Domain}, $Onto_{Domain}$);

5.2.3 $BN_{Problem}$ = $BN_{Problem}$ ∪ { BN_{Si} }; // add the new Bayesian network, BN_{Si}

5.3 return $BN_{Problem}$; // return the set of Bayesian Networks associated to all scenarios

6. KB_{BN} = *BN_KnowledgeAcquisition*($BN_{Problem}$); // derive probabilistic rules from each BN_{Si}

// build the knowledge base of the problem by knowledge integration and validation

7. $KB_{Problem}$ = *KnowledgeIntegration_Validation*(KB_{BN}, KB_{DM}, KB_{Domain}, $Onto_{Domain}$).

Some more details related to the proposed framework are given as follows.

A scenario is an instantiation of the problem.

The conceptual model, $CM_{Problem}$, contains only the key elements (key concepts) for solving the problem (being a sort of initial sketch), while the ontology, $Onto_{Domain}$ (i.e. with all sub-ontologies related e.g. to different scenarios) must include more terms (concepts and relations), being a detailed conceptual description of the problem domain, for all scenarios of the problem.

The environmental data set, DataSet (T_DB as denoted in Fig. 1), contains time series data (continuous data sets) and/or sample data (discrete data sets). Some of the data are measurements collected with monitoring stations (stationary or mobile) under a monitoring network (as e.g. an air quality monitoring network), other are samples collected at different locations and moments of time (as for example, in river water pollution analysis), while other data are characteristics of the environment area (e.g. area geographical characteristics, specific river characteristics).

The environmental data included in DataSet can contain measurements errors (e.g. from the automatic monitoring instrumentation), noise, uncertainty, missing data, redundancies, irrelevant information etc. Thus, some data cleaning and pre-processing techniques are performed at the beginning of step 4, before applying the inductive data mining algorithms. For example, some data visualization tools such as boxplots, time series plots, two-dimensional scatter plots or distributional plots can be applied in order to detect the presence of outliers, errors, missing values etc. Such problems can be solved (manually or automatically) by missing data imputation, filtering for noise removal or other solutions, usually, provided within the data mining software package that is used. Also, the cleaned and pre-processed data will be properly prepared for the application of the selected data mining algorithms in order to fulfill the required assumptions. Feature selection and principal component analysis are examples of pre-processing techniques performed for data preparation. Depending on the specific environmental problem (available data, analysis goal), it is decided which are the most suitable types of data pre-processing techniques required in step 4.

The domain ontology, $Onto_{Domain}$, is used for concepts and relations validation in steps 3, 4, 5 and 7, in order to select the right concepts and relationships between them. As during Bayesian Networks development (step 5) it is used the ontology (e.g. to define BN variables), it is not necessary to include it also in step 6, when it is performed probabilistic rules extraction from Bayesian Networks.

Starting from the environmental problem formulation, the following issues are defined:

1) a *set of environmental parameters* that need to be analyzed in order to solve the problem (environmental variables with their associated values - *numerical* or nominal, as e.g. measurements or observations and/or linguistic terms) and *a list of possible phenomena* associated to the environmental problem (linked to problem's scenarios),
2) *a data set for all/some environmental parameters* (available past and current data with specific site/area measurements, observations and characteristics),
3) *a goal variable* that need to be determined in order to solve the problem (as e.g. the occurring of a certain phenomenon, its possible effects on the environment, on ecosystems or on human activity, human settlements etc, and the decisions that should be taken in order to minimize the negative effects).

The set of the *environmental parameters* includes area or site-dependent parameters (e.g. pollution sources, specific water pollutants, and specific air pollutant), season-dependent parameters (which capture seasonality) and other parameters dependent on the problem. The human expert selects with higher priority the rules that include those parameters.

The core of the knowledge integration and validation step (step 7) is detailed as follows.

Suppose we have the following set of *m* rules from the three knowledge bases, KB_{BN}, KB_{DM}, KB_{Domain}, that need to be integrated into one knowledge base:

{ **IF** P_1 **THEN** Q_1; **IF** P_2 **THEN** Q_2; … ; **IF** P_m **THEN** Q_m }

Algorithm for Knowledge integration and validation (step 7)

1. **if** $Q_1 \neq Q_2 \neq \ldots \neq Q_m$ **then** // if rules have different conclusions
   1.1 **if** $P_1 = P_2 = \ldots = P_m$ **then** // rules have the same premises
      1.1.1 * the human expert checks the validity of each rule and selects the valid rules (VR);
      1.1.2 * include in $KB_{Problem}$ the rules from VR with the highest CNF and/or probability, which are not already included in $KB_{Problem}$;
   1.2 **else** // different premises
      // *case of rule chaining*
      1.2.1 **if** **some conclusions Q_j appear in the premises P_k of other rules* **then**
         i. * select the rules with the conclusions Q_j and the rules with the premises P_k;
         ii. * apply rule chaining for the selected rules and check the final result validity;
         iii. **if** * the final result is valid **then**
            * include in $KB_{Problem}$ the selected rules that are not already included;
      1.2.2 **else** // general case
         i. **if** * the rules are *final/end rules* (have the *goal* in conclusion) **then**
            * the human expert can include them in $KB_{Problem}$ with possible correction, and check rules' no redundancy;
         ii. **else if** *not final/end rules* (i.e. *isolated rules*) **then**
            * the human expert decides if they are eliminated or corrected;
2. **else if** $Q_1 = Q_2 = \ldots = Q_m = Q$ **then** // rules with the same conclusion Q
   2.1 * include in $KB_{Problem}$ the rule **IF** P_1 **OR** P_2 **OR** … **OR** P_m **THEN** Q, with human expert approval or correction, if it is not already included.

A rule is valid (step 1.1.1) if it is in accordance with the conceptual model, $CM_{Problem}$, or with the conceptual models of each identified scenario, CM_{Si} (i.e. the relation between the parameters from the rule's premise and conclusion is in accordance with the conceptual model, $CM_{Problem}$ or CM_{Si}).

The correction of a rule (step 1.2.2 i.) means the inclusion/exclusion of some additional parameter(s) in/from the rule's premise and/or conclusion in accordance with the conceptual models, $CM_{Problem}$ and CM_{Si}. Also, the rule correction can involve the change of its parameters' name or their symbolic values, in order to have uniformity in the final $KB_{Problem}$.

The final validation of the rules included in $KB_{Problem}$ it is performed by the human expert through reasoning, by rule chaining for all identified scenarios of the problem. During validation, the human expert checks the characteristics of the knowledge base: *coherence* (the conclusions provided by rule chaining are not contradictory if the initial facts are not contradictory), *no redundancy* (the rules are not repeated under different forms), and *completeness* (all useful rules for problem solving are included, i.e. the knowledge base allows complete resolution).

The data mining approach (rule induction) is used to discover rules in data sets (e.g. time series) and decision tables, and to identify the most correlated parameters, which are included as nodes in the Bayesian networks. In order to keep simplicity, several Bayesian networks with smaller complexity are built and analyzed. Probabilistic rules are derived from them. The problem domain ontology is used to design the Bayesian networks and to check rules consistency as well as to guide the selection of the right rules that are derived with inductive learning techniques and Bayesian networks. An important benefit of using inductive rule data mining in step 4 of KM-framework is that the knowledge derived from problem specific datasets are priority rules to the more general knowledge derived in step 3, being problem dependent. Actually, the rules from KB_{DM} are refined rules of those included in KB_{Domain}. Another advantage of using inductive learning techniques is their robustness to missing data (Cios et al., 2007).

The framework uses two types of inductive learning algorithms: decision tree based algorithms (indirect methods for rule generation from decision trees) and rule based algorithms (direct methods for rule generation directly from the data set). Depending on the environmental problems it is selected the proper inductive data mining algorithm by experimenting several ones from both types, for example, by choosing them from the algorithms available in the data mining software package (as e.g. Weka workbench (Witten et al., 2011), used for the case studies described in the next section, which has several types of inductive learning algorithms - decision tree and rule based). Examples of decision tree based algorithms that can be applied under the proposed framework are ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993) (implemented in Weka as J48 algorithm), M5 (Quinlan, 1992; Wang and Witten, 1997) (implemented in Weka as M5P algorithm), CART (Friedman, 1977; Breiman et al., 1984) and others (e.g. RP algorithms, REPTree etc). Some of these algorithms use information gain criteria (e.g. ID3, C4.5, REPTree) while others use least square criterion (e.g. M5, CART), mean absolute deviation criterion (e.g. CART) or other criteria (e.g. least absolute deviation). The rule based algorithms considered by the framework include CN2 (Clark and Niblett, 1989), M5Rules (the rule based version of M5, implemented in Weka), and other specific algorithms (as e.g. Decision Tables, PART, OneR that are implemented in Weka). Some of the inductive learning algorithms are sequential covering algorithms, learning one rule at a time (as e.g. CN2), while others (e.g. ID3, C4.5) are simultaneous covering algorithms, learning the entire data set simultaneously.

In the next section three case studies are presented, showing the framework application to different environmental problems in the domains of water, air and soil. Due to space limitation, a more detailed description is given only for the problem of navigable river resource management, presented in the first case study.

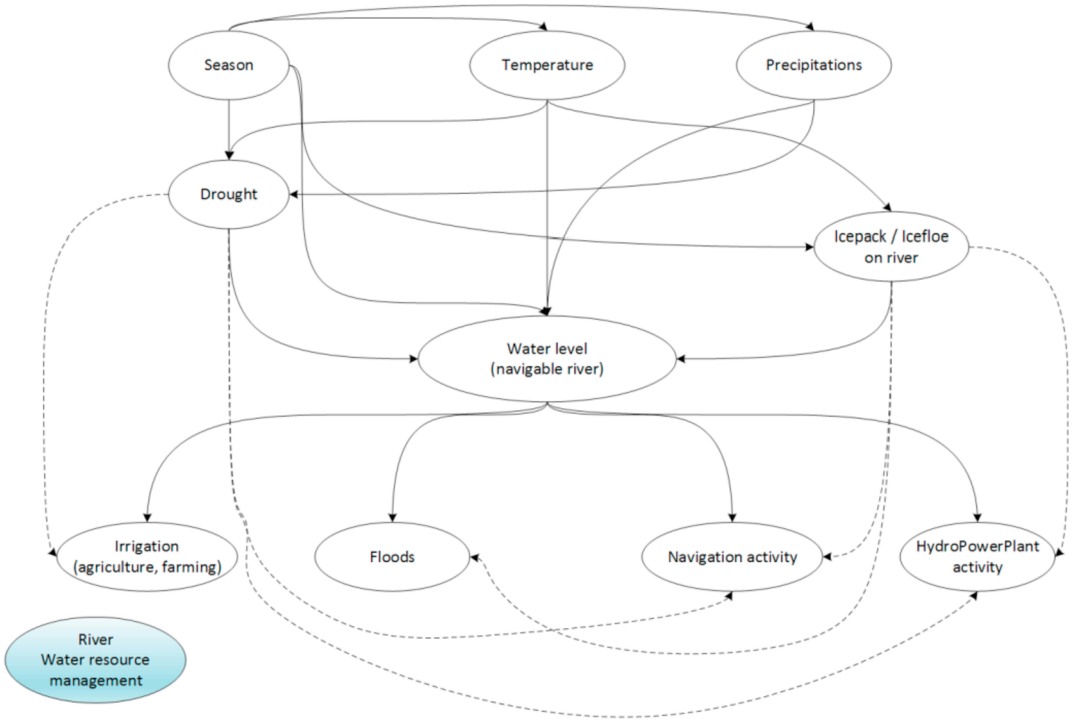

Fig. 2. The conceptual model for navigable river resource management problem (CM_{PW1}).

4. Case studies

4.1. Case study 1 – environmental domain: water

We are considering two problems from the water environmental domain: *water resource management* and *surface water pollution*. The application of the proposed framework is detailed as follows.

Problem (PW1): water resource management of a navigable river;

Problem description:

Water resource management is an environmental problem with major implications in life on earth (i.e. human life, flora and fauna), agriculture, industry etc. As an example we have considered the problem of water resource management for a navigable river, the Danube River. The main problems encountered in the river area due to water level variation are floods (occurring during rainfalls and ice rapid smelting), worsening or suspending navigation activity (during drought - usually, in the summer season, and icepack, during winter season), affecting hydropower plants activity and irrigation (during drought). In this case, the main goal of the water resource management is to avoid or to reduce the negative effects of each problem occurring in the river

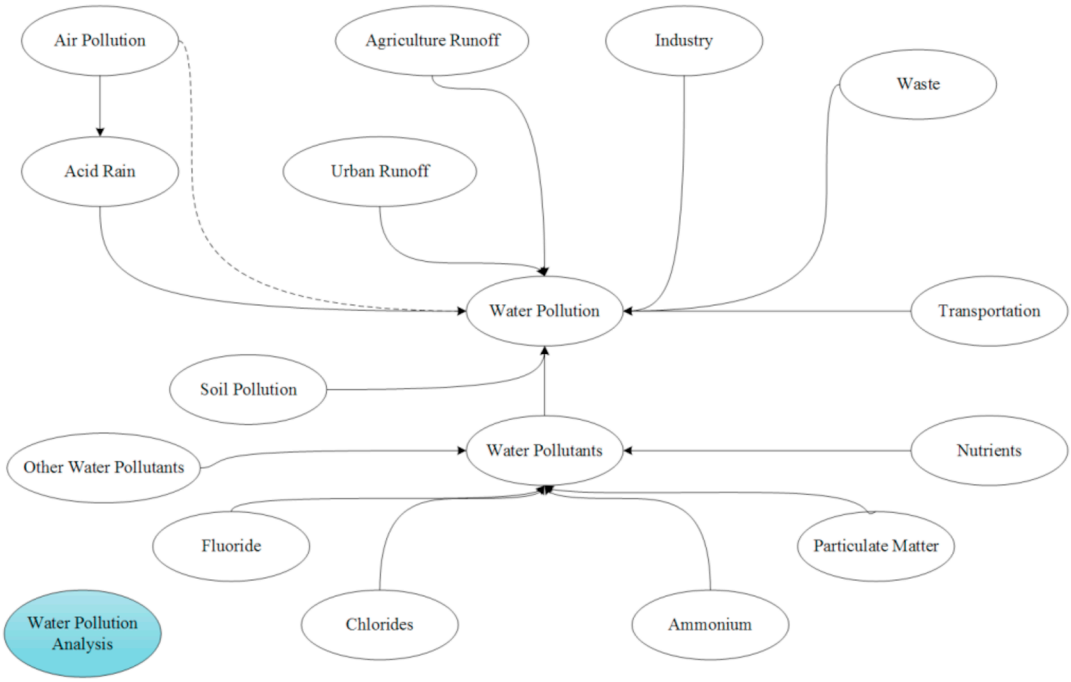

Fig. 3. The conceptual model for water pollution analysis problem (CM_{PW2}).

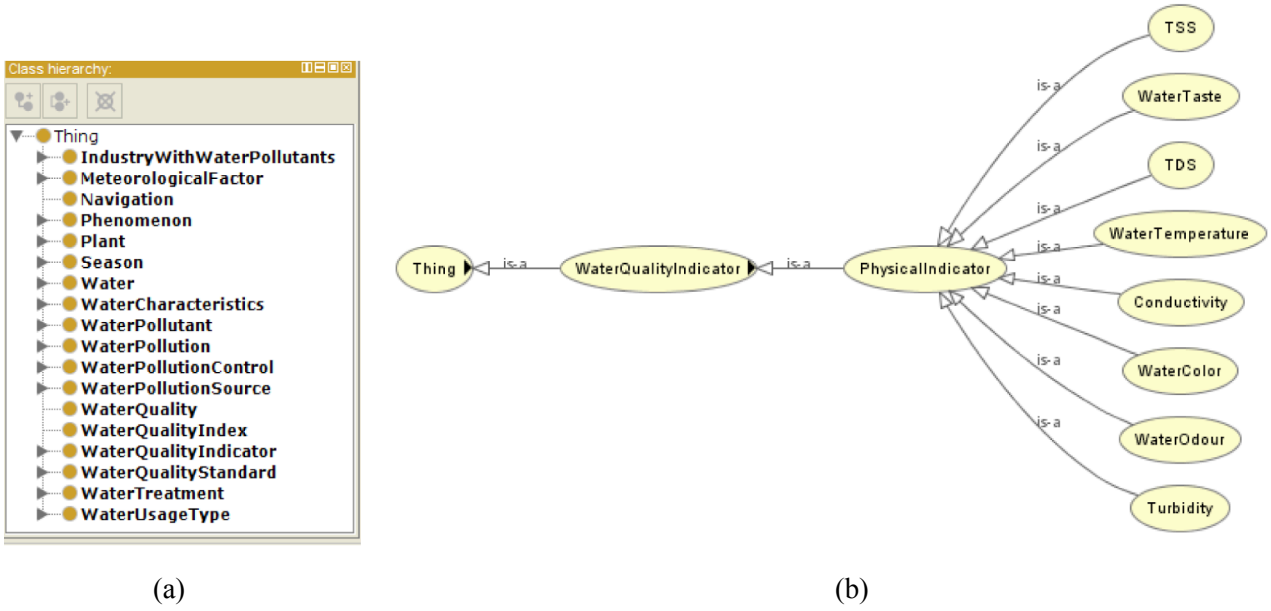

Fig. 4. (a) The class hierarchy of the Water-Onto ontology (OWL) in Protégé 4.3
(b) The PhysicalIndicator sub-class (Water-Onto ontology) in OWL Viz (Protégé 4.3).

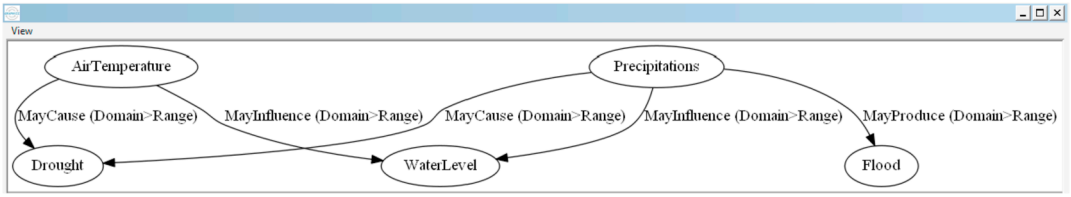

Fig. 5. Cause-effect graph for analyzing water level and two phenomena arising, drought and flood (in GraphViz 2.38).

area (the Danube River area), as for example, by informing the navigators (e.g. sending notification when a navigation problem occurs, due to drought or icepack forming), the local authorities (e.g. sending warning or alerts when floods can occur in a certain area or sending notification when drought will affect irrigation and thus, the agriculture and farming) or the hydropower plants management (e.g. sending notification when the river level is affected by drought).

Step 1. Conceptual model design (for PW1)

Starting from the problem description it was designed a conceptual model (CM_{PW1}) that is presented in Fig. 2. We have considered Water level (i.e. River level), Precipitations, Season, and Temperature, as the main parameters, Drought and Icepack, as the phenomena that can occur, Floods, as the potential effect of water level significant increase, IceRapidSmelting a potential effect in case a significant temperature increase is registered, and Irrigation, Navigation activity and HydroPowerPlant activity, as the affected activities due to drought, floods or icepack. The parameters set, and the corresponding sets for the identified phenomena, effects and affected activities are given below.

Parameters Set = {WaterLevel, Temperature, Precipitations, Season};
Phenomena = {Icepack, Drought}; *Effects* = {Floods, IceRapid Smelting};
AffectedActivities = {Irrigation, NavigationActivity, HydroPower PlantActivity}.
Problem (PW2): surface water pollution analysis;
Problem description:
Surface water pollution analysis (e.g. for rivers, lakes) is another important environmental problem with implications on aquatic life, human health etc. In this case, the goal is to determine the level of

Table 2
PW1 problem related phenomena's causes and affected activities.

Phenomena	Causes	Affected activities/ entities
Flood	*Rainfall* (in any season, normally, except winter), *Ice rapid smelting* (when high/very high temperatures are registered, normally, at the end of winter or begin of spring);	- Agriculture - Farming - Human settlements
Drought	*Lack of precipitations* and *very high temperature* (normally, in summer and possible, in autumn);	- River Navigation - Hydropower plants - Agriculture (Irrigation) - Farming
Icepack, Ice bridge	*Low* or *very low temperature* of the river water (normally, in winter and more rarely at begin of spring).	- River Navigation - Hydropower plants

surface water pollution. The main sources of surface water pollution are agriculture runoff, urban runoff, industry, transportation, waste, soil pollution, and air pollution. Examples of water pollutants are nutrients, fluoride, chlorides, ammonium, and particulate matter.

Step 1. Conceptual model design (for PW2)

Starting from the problem description we have designed a conceptual model (CM_{PW2}) that is shown in Fig. 3. The main parameters that were considered are Water Pollution Source and Water Pollutant, while the goal variable is Water Pollution Level.

Parameters *Set* = {WaterPollutionSource, WaterPollutant}; *Goal* = {WaterPollutionLevel}.

Table 3
List of parameters associated to the PW1 problem (name, domain, abbreviated names, set of linguistic values).

PW1 Parameter name	Domain [measurement unit]	Abbreviated name	Set of linguistic terms
Season			{autumn, winter, spring, summer}
Precipitations	numeric [mm]	PS	{significant, insignificant}
AirTemperature	numeric [^{0}C]	AirT	{very high, high, normal, low, very low}
AirTemperatureVariation	numeric [^{0}C]	AirTVar	{increase, stationary, decrease}
WaterTemperature	numeric [^{0}C]	WaterT	{very high, high, normal, low, very low}
WaterTemperatureVariation	numeric [^{0}C]	WaterTVar	{increase, stationary, decrease}
Weather (last days weather)			{ warm, cold, snowfall, rainy, droughty, Strong Drought, Light Drought }
Phenomenon			{Drought, Light Drought, Possible Icepack, IcePack, IceBridge, None}
RiverLevel	numeric [cm]	RiverLS	{low, normal, high, very high}
RiverLevelVariation	numeric [cm]	RiverLVar	{slight increase, increase, strong increase, stationary, decrease}
IceRapidSmelting			{True, False}
Goal variable 1: *Flood*			{No, Possible, FloodAlert}
Goal variable 2: *Decision*		Dec	{No, Inf, Res, Sto}

Step 2. Ontology development (for both problems, PW1 and PW2)

We have developed one prototype ontology for the water domain, $Onto_{Water}$ as denoted in the framework, named Water-Onto, for both problems, PW1 and PW2, starting from the two conceptual models, CM_{PW1} and CM_{PW2}. The main concepts related to water resource management for a navigable river and water pollution were included as classes. The ontology implementation was done in Protégé 4.3 as an OWL ontology. Fig. 4 shows the class hierarchy of Water-Onto and a screenshot with concepts of the PhysicalIndicator sub-class as e.g. Total Suspended Solids (TSS), Conductivity, Turbidity, WaterTemperature.

Apart from the concepts that were included in the Water-Onto ontology, some problem specific relations between concepts were defined as object properties. Examples of such relations are *MayInfluence*, *MayProduce*, *MayCause*. The default relations are the taxonomic ones, *is-a* and *ako* (a kind of). Fig. 5 shows a cause-effect graph for the PW1 problem. Air temperature *may influence* water level through evaporation (usually, during summer, when very high temperature can be registered), and *may cause* drought. The quantity of precipitations *may influence* the water level, *may produce* flood (when important quantity of precipitations are registered) or *may cause* drought (by lack of precipitations).

Step 3. Knowledge Acquisition (for both problems, PW1 and PW2)

In this step, the main knowledge related to each problem solving (PW1 and PW2) was taken from literature and human experts according to the designed conceptual models (CM_{PW1} and CM_{PW2}) and was described as IF-THEN rules and decision tables, being included in the corresponding knowledge base (i.e. $KB_{PW1\text{-}Domain}$ for PW1 and $KB_{PW2\text{-}Domain}$ for PW2). The Water-Onto ontology was used to select the proper concepts included in knowledge representation (e.g. parameters' names). The knowledge uncertainty was quantified by linguistic terms (fuzzy model, in the case of decision tables) and confidence factors (in the case of rules).

Some examples of knowledge and their source are given below for PW1 problem.

Problem PW1 − $KB_{PW1\text{-}Domain}$:

Table 2 synthesized human expert knowledge related to the main causes of the three phenomena that were considered in PW1 problem, *flood*, *drought*, *icepack* and *ice bridge* occurrence, and the affected activities or entities to whom the river resource management authority will send notifications. In case of agriculture, farming and human settlements, local authorities will be informed, while in case of river navigation, periodically reports will be sent to river navigators.

Table 3 shows the names of the parameters that were considered in solving the PW1 problem, their domain and measurement unit, the abbreviated name and their associated set of linguistic terms.

The first goal variable, Flood refers to the possibility of flood occurrence. The second goal variable, *Decision*, refers to the navigable river resource management decision and can represents: *no decision* (no problem occurred), *send an informing notification* (Inf), *send a notification for restricted activity* (Res) *or for stopping activity* (Sto).

We have considered the following site/area dependent river level's thresholds: River Warning Level (i.e. the river level above which it is possible to occur flood), River Flood Level (i.e. the river level above which the flood occurs), and River Minimum Level (i.e. the level of the river under which some activities will be affected as e.g. navigation, irrigation, hydropower plants activity), with the corresponding values for a certain locality along the river, where the Water Level is measured.

Examples of *rules related to site/area dependent parameters* (for setting linguistic terms, as e.g. for the river level variable, RiverLS, and the values of other variables, Flood, Decision):

```
Rule PW1-S1
IF WaterLevel < RiverWarningLevel AND WaterLevel >= RiverMinimumLevel THEN
    RiverLS = normal
    Flood = No CNF 100
    Decision = Inform; // normal activity for navigation, hydropower plants etc

Rules PW1-S2
IF WaterLevel >= RiverWarningLevel AND WaterLevel < RiverFloodLevel THEN
    RiverLS = high
    Flood = Possible CNF 60
    Decision = Informing "Flood warning"; // informing local authorities, navigators etc

Rule PW1-S3
IF WaterLevel >= RiverFloodLevel THEN
    RiverLS = very_high
    Flood = FloodAlert CNF 100
    Decision = Informing "Flood alert!"; // informing local authorities, navigators etc

Rule PW1-S4
IF WaterLevel < RiverMinimumLevel THEN
    RiverLS = low
    Decision = Stopped // notification to navigators about RiverMinimumLevel alert;
```

Another set of rules are referring to the setting of linguistic terms associated to the numerical parameters set. Some examples are given below. The estimation of the significant, normal and insignificant level of precipitations during a certain season and geographical area is established by human experts. The PS parameter denotes the linguistic term associated to a certain numerical value of the Precipitations parameter. We have to notice that each of the following three rules is established for each season. The same for temperature (air or water), where we have the linguistic values of *very high*, *high*, *normal*, *low*, *very low*, which depends on the season. As we refer to a geographical area with temperate climate, the Season parameter can have a value from the following set: {spring, summer, autumn, winter}.

Rule PW1-P1

IF Precipitations < $PrecipMinLevel^{Season}$ THEN

PS = insignificant;

Rule PW1-P2

IF Precipitations >= $PrecipMinLevel^{Season}$ AND Precipitations <= $PrecipMaxLevel^{Season}$

THEN

PS = normal;

Rule PW1-P3

IF Precipitations > $PrecipMaxLevel^{Season}$ THEN

PS = significant;

Table 4
-a. PW1-DT1 Decision Table for the autumn season (selection).

PS	AirT	AirTVar	WaterT	WaterTVar	Weather	Phenomenon	RiverLS	RiverLVar	Flood	Dec
insignificant	high	stationary	high	increase	Light Drought	Light Drought	normal	decrease	No	No
insignificant	high	stationary	high	increase	warm	None	normal	stationary	No	No
significant	normal	stationary	normal	stationary	rainy	None	normal	increase	Possible	No
significant	low	decrease	low	decrease	rainy	None	high	increase	Possible	No
insignificant	low	increase	low	increase	cold	None	normal	stationary	No	No
insignificant	high	increase	high	increase	Strong Drought	Drought	low	decrease	No	Res
significant	very low	decrease	very low	decrease	Snowfall	Possible Icepack	normal	Slight increase	No	Res
significant	normal	stationary	normal	increase	Rainy	None	very high	Strong increase	Possible	Inf
insignificant	normal	stationary	normal	stationary	Rainy	None	normal	Slight increase	No	No
insignificant	very low	decrease	very low	decrease	Droughty	Possible Icepack	low	stationary	No	Res
significant	low	decrease	low	decrease	Snowfall	Icebridge	normal	Slight increase	No	Sto
significant	low	increase	low	increase	Snowfall	Icepack	high	increase	No	Res
insignificant	high	increase	high	increase	Warm	Icepack	high	increase	Possible	Res
insignificant	very low	stationary	low	stationary	Very cold	Icebridge	normal	stationary	No	Sto
insignificant	low	decrease	low	stationary	Cold	Icepack	normal	stationary	No	Res
significant	normal	increase	normal	stationary	Snowfall	Icepack	normal	increase	No	Res
normal	normal	increase	normal	increase	Rainy	Icepack rapid smelting	high	increase	Possible	Inf
significant	high	increase	high	increase	Rainy	Icepack rapid smelting	very high	increase	Flood Alert	Inf
significant	normal	increase	normal	increase	Droughty	Light Drought	low	increase	No	No
insignificant	high	increase	high	increase	Droughty	Drought	low	decrease	No	Res
significant	normal	increase	normal	stationary	Rainy	None	high	increase	Possible	Inf
significant	high	stationary	high	stationary	Rainy	None	very high	increase	Flood Alert	Inf
insignificant	normal	increase	normal	increase	Warm	None	normal	decrease	No	No
insignificant	low	decrease	low	decrease	Rainy	None	low	Slight increase	No	No
significant	high	increase	high	increase	Warm	None	normal	increase	No	No
insignificant	very high	increase	high	increase	Droughty	Drought	very low	Strong decrease	No	Sto
insignificant	high	increase	high	increase	Droughty	Drought	low	decrease	No	Res
significant	high	stationary	high	increase	Rainy	None	high	increase	Possible	Inf
significant	normal	stationary	normal	increase	Rainy	None	very high	Strong increase	Flood Alert	Inf

Table 5
PW1-DT2 decision table (selection).

Phenomenon	RiverLS	RiverLVar	NavigationActivity	HydroPowerPlantActivity
Light Drought	normal	decrease	Normal (CNF 80)	Normal (CNF 80)
None	normal	increase	Normal	Normal
None	normal	stationary	Normal	Normal
Drought	low	decrease	Restricted	Restricted
Possible Icepack	normal	stationary	Possible Restricted	Possible Restricted
Possible Icepack	low	stationary	Possible Restricted	Possible Restricted
Icepack, Icebridge	normal	stationary	Stopped	Stopped
Icepack	high	increase	Restricted	Possible Restricted
Drought	low	increase	Restricted	Possible Restricted
None	high	increase	Normal	Normal

Decision tables are built by human expert starting from the knowledge related to PW1 problem analysis and already included in the conceptual model (CM_{PW1}), as for example:

Water level is influenced mainly by the precipitations quantity and evaporation. However, during winter when icepack are formed on the river, and when higher temperature are recorded, possible floods can appear (due to rapid icepack smelting). Rainfalls can increase significantly the level of the river, and thus, can produce floods affecting hydropower plant activity, agriculture and human settlements. Evaporation can influence water level, usually, during summer when higher values of air temperature are registered and also, few precipitations or lack of precipitation are recorded.

Table 4-a, b, c, d present selected examples from PW1-DT1 Decision Table for each season of a temperate climate (autumn, winter, spring, summer), in which it is analyzed the possibility of occurring the phenomena of Icepack, Ice bridge, Drought or Flood.

Table 5 presents some examples from the PW1-DT2 decision table, in which are explicitly analyzed the possible effects of the drought and icepack phenomena on the activity of hydropower plants and navigation. The parameters that were used are Phenomenon, RiverLS, RiverLVar, NavigationActivity and HydroPowerPlantActivity. The linguistic terms used for the last two parameters are the following: Normal, Possible Restricted, Restricted, Stopped. We have specified a confidence factor, CNF, for the Normal value in the first example. The default value of CNF for the other examples is 100.

A set of heuristic rules can be extracted from the rows of the previous decision tables. However, we shall see in step 4 (data mining) that a smaller number of rules can be extracted with inductive machine learning techniques, keeping only the most representative (correlated) parameters. Other examples of heuristic rules with confidence factors established by human experts are given below.

Rule PW1-A1

IF PS = Significant AND Season = Spring THEN RiverLevel = High CNF 95;

Rule PW1-A2

IF PS = Insignificant AND Season = Summer THEN RiverLevel = Low CNF 97;

Rule PW1-A3

IF PS = Significant AND Season = Autumn THEN RiverLevel = High CNF 80;

Rule PW1-A4

IF PS = Insignificant AND Season = Autumn THEN RiverLevel = Low CNF 87;

Rule PW1-A5

IF PS = Significant AND Season = Winter THEN RiverLevel = High CNF 75;

Step 4. Knowledge discovery - via data mining (for PW1 problem)

During the knowledge discovery step of the framework (i.e. step 4) are performed data cleaning, data pre-processing and inductive data mining.

We have used public available datasets (collected from hydro-meteorological reports) provided by the Down-Stream Danube River Administration R.A. Galati (Romania) www.afdj.ro/to/cotele-dunarii, and the decision tables generated in the previous step.

The main data included in a hydro-meteorological report are: air temperature at 7 a.m., the water temperature (daily average value), the river level, the river level variation and trend, the specification of season specific phenomena (e.g. snow level height, presence of icepack or ice bridge), other meteorological data (such as precipitations, wind speed, wind direction, atmospheric pressure) etc. These values are given for each locality (observation point) situated along the Dunarea River. However, we have encountered a problem, precipitations quantity missing from the collected reports, which was compensated by the decision tables given by human experts.

We have used the Weka 3.8.0 data mining software package. Before applying the data mining algorithms, the data sets were cleaned, checked and pre-processed by using the available tools from Weka (as e.g. graphical visualization tool – plot matrix visualizing all environmental variables, filters for noise removal). Also, some pre-processing steps required by specific inductive data mining algorithms were performed (as e.g. normalization, feature selection). Two types of inductive learning techniques implemented in Weka, decision trees and rule algorithms, were applied to decision tables and to the cleaned and pre-processed datasets in order to derive a set of heuristic rules and the most correlated parameters. Several experiments with different classifiers from the two type of inductive learning techniques were performed under the test mode 10-fold cross validation, and the best results were registered. The specific inductive learning algorithms that were selected in this case study are M5P, REPTree and J48 from the decision tree based algorithms and M5Rules from the rule based algorithms. We give a synthesis of some experiments that were performed providing examples of rules or correlations that were derived and included in KB_{DM} knowledge base. Some details related to the data pre-processing and inductive data mining steps are given for all experiments except the one performed on decision tables.

Experiment 1 Identify the correlations between water level, air temperature, water level variation, and icepack presence during winter season (January–February 2017).

Dataset: Num-RiverDunarea2017-Jan-Feb.arff (numerical values from hydro-meteorological reports).

Attributes: AirTemperature, WaterTemperature, RiverLevel, RiverLevelVariation, IcepackPresence.

A1. Data Pre-processing:

1) Attribute selection with *best first* strategy:
   Selected attributes: WaterTemperature, RiverLevel, RiverLevelVariation;
2) Data Normalization.

A2. Data Mining - Inductive learning (Inductive DM algorithms: M5Rules, M5P, REPTree)

Best results: M5Rules classifier results (validated by human experts):
RiverLevel depends on WaterTemperature and IcepackPresence.

Experiment 2 Identify the correlations between water level, air temperature, water level variation and phenomenon presence in any season.

Dataset: Nom-***RiverDunarea2017***. arff (numerical and nominal values).

Attributes: ***AirTemperature***, WaterTemperature, RiverLevel, RiverLevelVariation, Phenomenon, Season.

A1. Data Pre-processing:

1) Attribute selection with *best first* strategy:
   Selected attributes: WaterTemperature, RiverLevel, RiverLevelVariation;
2) Data Normalization.

A2. Data Mining - Inductive learning (Inductive DM algorithms: M5Rules, M5P, REPTree)

Best results: M5Rules classifier results (validated by human experts):
RiverLevel depends on WaterTemperature and Phenomenon (drought, icepack etc).

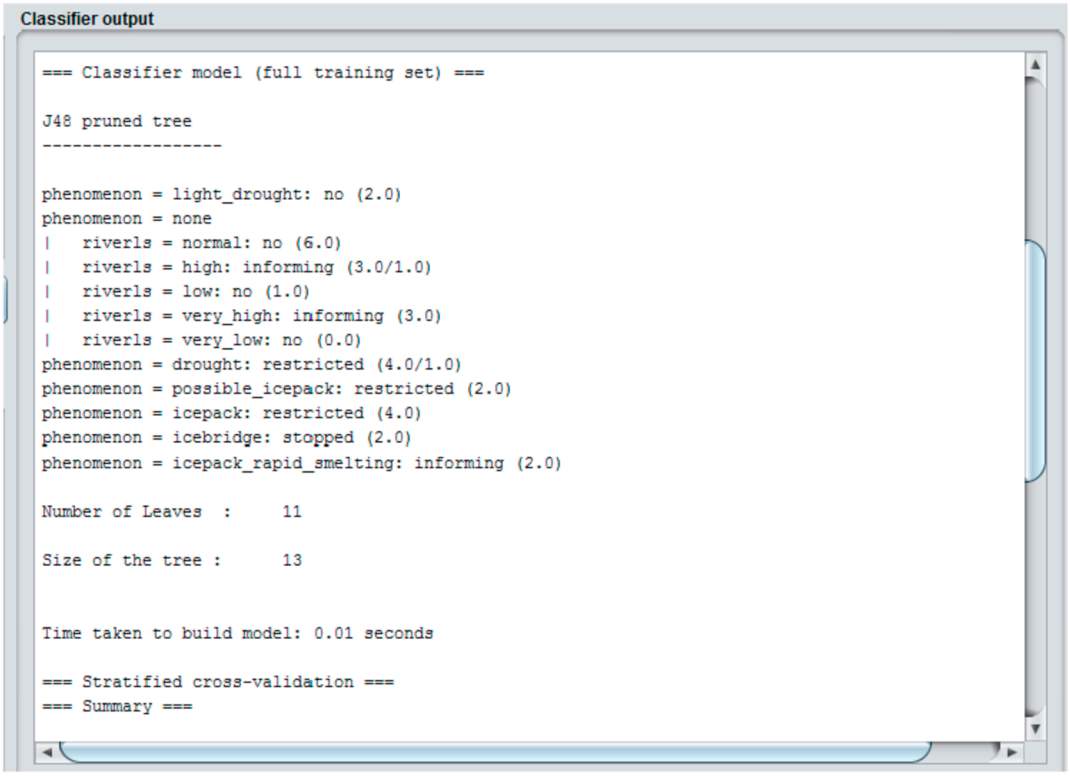

Fig. 6. Screenshot with J48 pruned tree (in Weka 3.8.0).

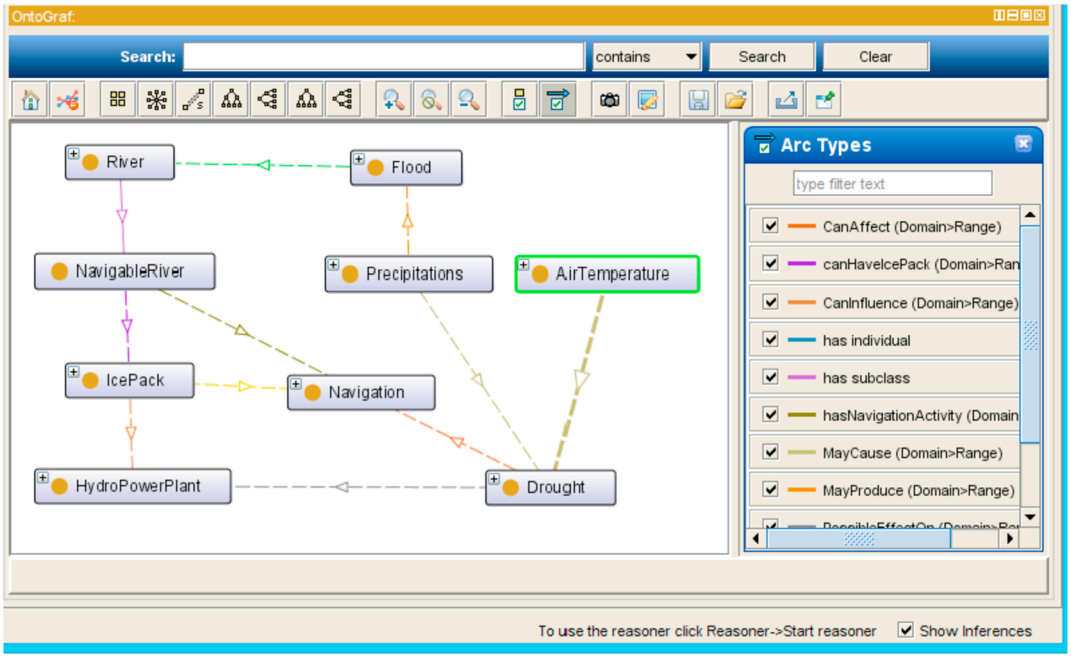

Fig. 7. Water-Onto *drought-flood-icepack combined scenario sub-ontology* graph (in OntoGraf, Protégé 4.3).

Thus, from the two experiments it was derived that AirTemperature parameter is not relevant for our problem, and it can be ignored in further rules. Actually, WaterTemperature parameter is important. Also, the RiverLevel parameter depends on WaterTemperature and Phenomenon. The importance of the Phenomenon parameter to the RiverLevel parameter evolution was used in the next step of the framework, for Bayesian network design.

Experiment 3 Identifying the most representative rules from decision tables of KBDomain. Examples of rules mined from the PW1-DT1 decision tables with J48 classifier (C4.5 like algorithm), which gave the best results, are shown in Fig. 6. The results for the statistic parameters are: Kappa statistic = 0.7119, mean absolute error (MAE) = 0.1212,

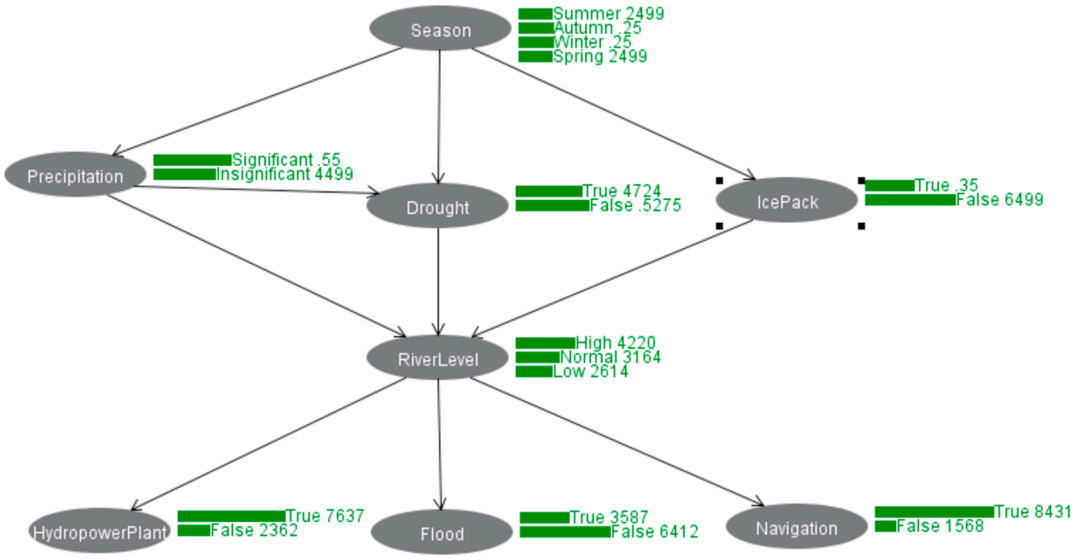

Fig. 8. The Bayesian network for *drought-flood-icepack* combined scenario (Weka 3.8.0).

and root mean squared error (RMSE) = 0.286.

Rule PW1-DM-1

IF Phenomenon = none AND RiverLS = high THEN Decision = Informing.

Rule PW1-DM-2

IF Phenomenon = icepack THEN Decision = Restricted.

Rule PW1-DM-3

IF Phenomenon = icebridge THEN Decision = Stopped.

Some of the rules from $KB_{PW1\text{-}DM}$ are used for building the Bayesian networks (e.g. when setting the values from the conditional probabilities tables).

Step 5. Bayesian network development

PW1 Problem Scenarios:

Three scenarios were identified for PW1 problem: drought occurrence and its effects, icepack and ice bridge occurrence and their effects, flood occurrence and its effects. We have combined all three scenarios into one scenario (the *drought-flood-icepack combined scenario*), which is shown in Fig. 7 as a graph (in OntoGraf, Protégé 4.3).

Starting from combined scenario, we have built in Weka the Bayesian network shown in Fig. 8. The BN nodes were set by using the Water-Onto ontology and the relations between them as well as the Conditional Probability Tables (CPT) were set according to the rules derived in the previous two steps of the framework. We have to notice that in contrast with the scenario shown in Fig. 7, we have eliminated AirTemperature from the Bayesian network as it was less relevant, for all three cases and we have included the RiverLevel parameter which is more important in two cases (drought and flood), ignoring the case of IcePackRapid smelting which can be managed by rules from KB_{Domain}.

Step 6. Knowledge acquisition from Bayesian networks

From the Bayesian network developed at the previous step we have derived the probabilistic rules. Examples of such rules are given below.

Rule PW1-BN7

IF RiverLevel = high $P(0.42)$ THEN

HydropowerPlant = True $P(0.76)$; // i.e. active

Flood = True $P(0.35)$; // i.e. flood occurrence

Rule PW1-BN2

IF RiverLevel = low P(0.26) THEN

Navigation = false $P(0.15)$; // i.e. restricted or stopped

Rule PW1-BN3

IF RiverLevel = normal THEN

Navigation = true $P(0.84)$;

where, P () is used to denote the probability.

Step 7. Knowledge integration and validation.

The knowledge derived in the previous steps are integrated into the final knowledge base KB_{PW1} according to the algorithm described in section 3 for step 7 of KM-Framework. Repeated rules are eliminated, while other rules are corrected, as for example, in case of the probabilistic rules, the name of the parameters and some symbolic values are changed, in order to keep rules' uniformity in KB_{PW1}. Also, new rules could be added in order to cover new problem's scenarios. The validation of the knowledge base was performed on several scenarios of the problem, via rule chaining and final result analysis by the human experts.

We present rule chaining in three particular cases of the PW1 problem.

Case 1. (winter phenomenon – ice bridge on Dunarea River):

Initial Facts:

Observation point: Giurgiu (Km 493); Date: February 2, 2017; Hour: 8 am;

RiverLevel = 187+11 cm; RiverLevelVariation = -(1-3) cm (decrease)

WaterTemp = 0.1 ^{0}C; AirTemp = -11.9 ^{0}C; IceBridge (Km 479-480)

Applied Rules:

Rule PW1-S1 (*Deduced conclusions*: RiverLS = normal, Decision = Informing),

Rule PW1-DM-3 (*Deduced conclusions*: Decision = Stopped),

Rule PW1-BN3 (*Deduced conclusions*: Navigation = true *P*(0.84))

Goal variable: Decision = Stopped the NavigationActivity.

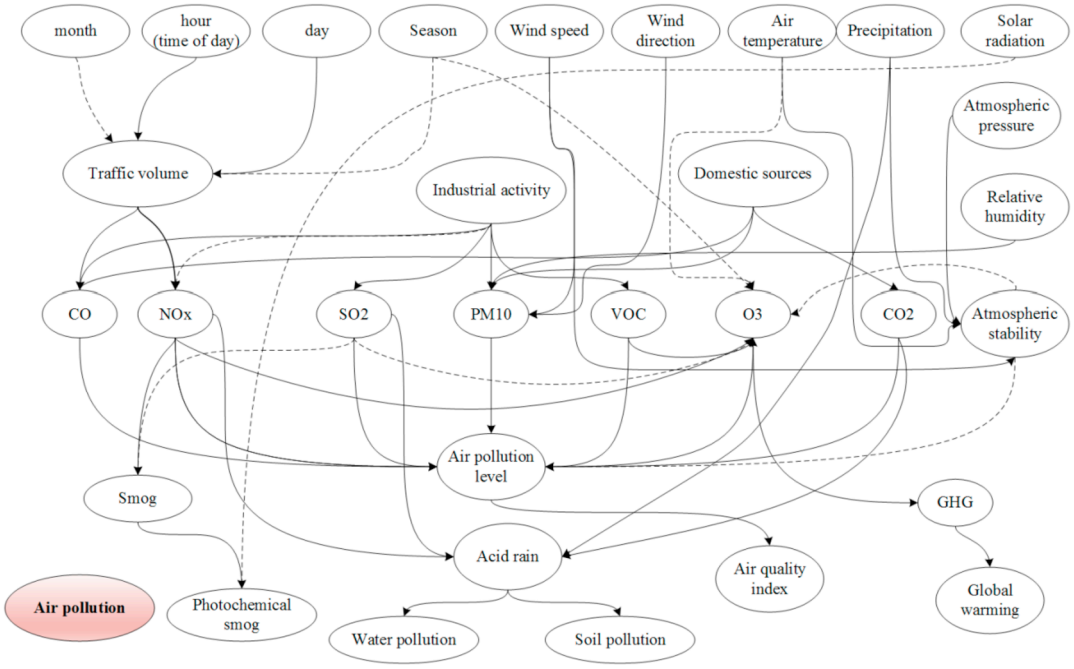

Fig. 9. The conceptual model of air pollution problem (CM_{PA1}).

In this case, the conclusion of the probabilistic rule, PW1-BN3 is kept together with the conclusion of rule PW1-DM-3, and finally, the last one conclusion is chosen as true in the given scenario context. This simple case provided an idea about the problems that should be tackled by the knowledge integration algorithm proposed in section 3.

Case 2. (drought scenario)

Initial Facts:

Observation point: Giurgiu (Km 493); Date: July 7, 2017; Hour: 8 am;

RiverLevel = 33 cm; RiverLevelVariation = -9 cm (decrease)

WaterTemp = 26.8 ^{0}C; Phenomenon = Drought

Applied Rules:

Rule PW1-S4 (*Deduced conclusions*: RiverLS = low, Decision = Stopped),

Rule PW1-DM-7 (*Deduced conclusions*: Decision = Restricted),

Rule PW1-BN2 (*Deduced conclusions*: Navigation = false *P*(0.15)

Goal variable: Decision = Restricted;

In this case, the conclusion of the probabilistic rule, PW1-DM-7 is kept as true in the given scenario context.

Case 3. (ice rapid smelting scenario - new scenario)

Initial Facts:

Season = spring; AirTVar = increase; Phenomenon = Icepack; RiverLS = high; For this case we have added the following rules during rules integration:

Rule PW1-P26
IF Phenomenon = Icepack AND AirTVar = strong_increase
THEN IceRapidSmelting = True CNF 90;

// check RiverLevel value with rules PW1-Si comparing with the limits specific to the observation point and set RiverLS

Rule PW1-P27
IF IceRapidSmelting = True AND RiverLS = high
THEN Flood = FloodAlert;

Rule PW1-P28
IF IceRapidSmelting = True AND RiverLS = normal
THEN Flood = PossibleFlood CNF 80;

Rule PW1-N1
IF Flood = FloodAlert THEN
Notification = Informing; // all entities: local authorities, navigators and power plants management;

Rule PW1-BN7
IF RiverLevel = high $P(0.42)$ THEN
HydropowerPlantActivity = normal $P(0.76)$;
Flood = FloodAlert $P(0.35)$;

The last rule was corrected, i.e. the name of the parameters and some symbolic values were changed in order to keep rules' uniformity in KBPW1.

4.2. Case study 2 – environmental domain: air

We are considering two problems from the air environmental domain: *air pollution analysis* (PA1) and *air pollution short-term prediction (ozone prediction)* (PA2). The application of the proposed framework is detailed as follows.

Problem (PA1): Air pollution analysis;

Problem description:

We have considered the air pollution analysis problem. The main air pollutants are CO, NO_x, SO_2, PM, O_3, VOC, which are usually monitored in most countries, under national air quality monitoring networks. The air pollution degree is influenced apart from the air pollutants concentration levels by other factors such as meteorological ones (e.g. wind speed, wind direction, precipitation, relative humidity, temperature, solar radiation, atmospheric pressure etc), season, time etc. The main sources of air pollution are transportation (traffic), industrial activities and domestic sources (especially, heating). Air pollution episodes can have potential negative effects on human health, being an important environmental issue that should be properly managed for increasing the quality of life in urban areas. The current legislation in most worldwide countries imposed limits for the concentration of major air pollutants and population warning via public available information of air quality

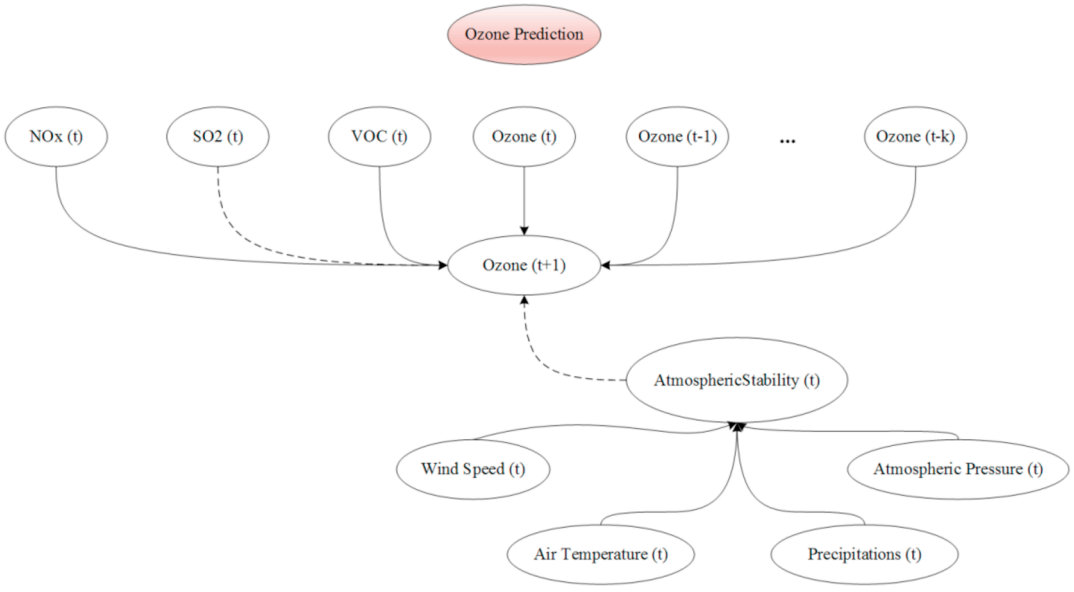

Fig. 10. The conceptual model of the next hour ozone prediction problem (CM_{PA2}).

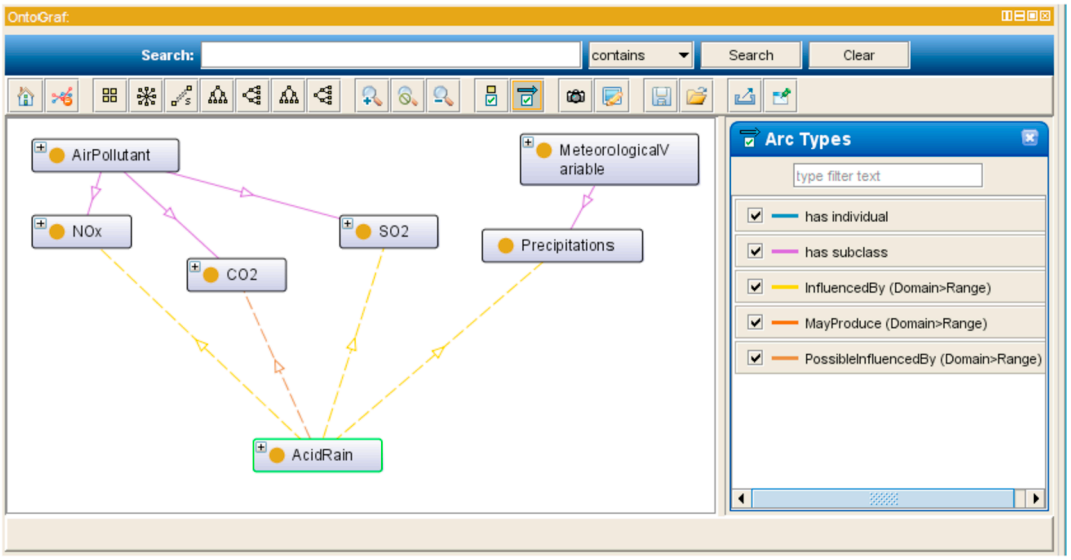

Fig. 11. AirPollution-Onto-1 *acid rain sub-ontology* graph (in OntoGraf).

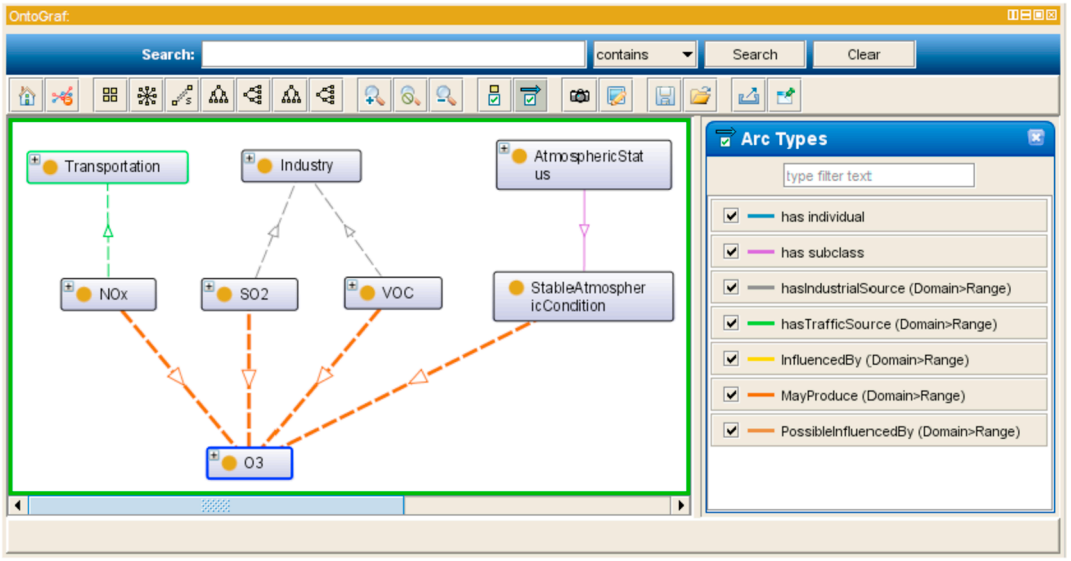

Fig. 12. AirPollution-Onto-1 *ozone pollution sub-ontology* graph (in OntoGraf).

index (general and specific to each air pollutant) in polluted areas (e.g. urban, sub-urban) with color codes related to possible human health effects and risks, and recommendation to reduce them. One of the air pollution phenomenon that can affect the quality of water and soil is acid rain, which can appear during rains and air pollution episodes due to some air pollutants (as e.g. NO_x, SO_2, CO_2) higher concentrations. Another air pollution problem is ozone appearance (especially, ground level ozone affecting soil quality), which is derived from chemical reactions determined by NO_x, VOC, under certain meteorological conditions (e.g. light wind, warm temperature, clear sky, i.e. stable atmospheric conditions), and is increased by SO_2 presence.

Step 1. Conceptual model design (for PA1)

Starting from description of the PA1 problem, we have designed a conceptual model for air pollution analysis, CM_{PA1}, which is presented in Fig. 9.

Problem (PA2): Air $_{pollution}$ short-term prediction (ozone);
Problem description:

Table 7
PA2-DT1 Decision Table for the next hour ozone prediction scenario (selection).

NOx	SO2	StableAtmCond	OzoneTrend (last days)	Next day Ozone
high	high	stable	increase	higher (increase)
low	low	stable	stationary	normal (stationary)
high	low	stable	increase	higher
low	high	stable	increase	higher
low	low	instable	stationary	lower

The second problem is ozone prediction by using current and past values of ozone concentration (*k* hours ago) and current concentration of some air pollutants that can influence ozone concentration level, as NO_x, SO_2 and VOC.

Step 1. Conceptual model design (for PA2)

Fig. 10 shows a conceptual model for next hour ozone prediction problem (CM_{PA2}).

Table 6
PA1-DT1 Decision Table for the acid rain scenario (selection).

NOx	SO2	CO2	Precip	Season	AcidRain	WaterPollution	SoilPollution
low	low	low	insignificant	any	no	no	no
low	low	high	significant	Autumn, Winter	Possible CNF 75	Possible CNF 70	Possible CNf 75
low	high	high	significant	Autumn, Winter	Possible CNF 85	Yes CNF 80	Yes CNF 85
high	high	high	significant	Autumn, Winter	Yes CNF 98	Yes CNF 90	Yes CNF 98
high	high	low	significant	Spring, Summer	Yes CNF 95	Yes CNF 90	Yes CNF 90
high	low	low	significant	Spring, Summer	Yes CNF 80	Yes CNF 75	Yes CNF 75
high	high	high	insignificant	Spring, Summer	No	No	No
low	high	low	insignificant	Spring, Summer	Yes CNF 80	Yes CNF 75	Yes CNF 75
low	low	high	insignificant	Winter	No	No	No

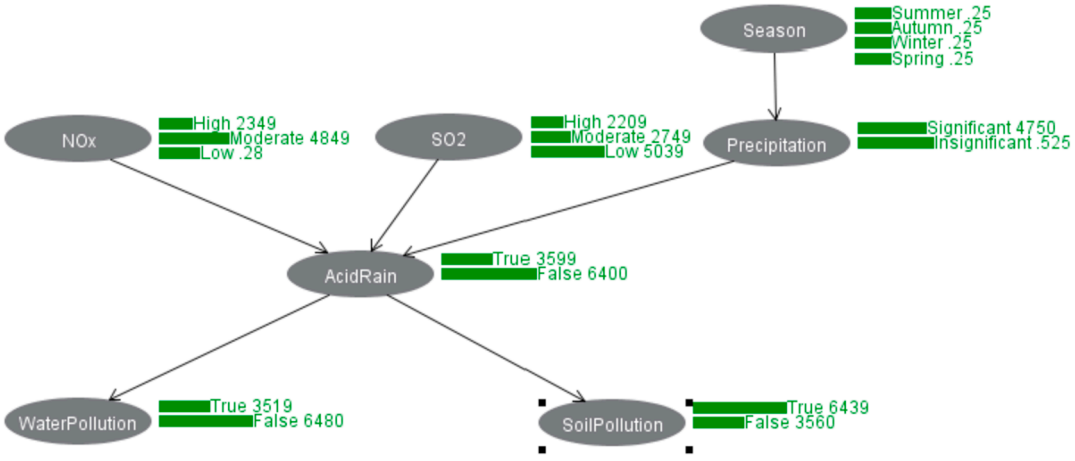

Fig. 13. The Bayesian network for Acid Rain scenario – PA1 (Weka 3.8.0).

Step 2. Ontology development (for both problems, PA1 and PA2)

Starting from the conceptual model we have identified the main concepts and relations between them and we have developed an ontology for the air pollution domain, AirPollution-Onto-1, implemented in Protégé 4.3 in owl format. The current version of the ontology has 143 classes, 29 object properties, 22 data properties, 319 logical axioms, 588 axioms, 43 individuals. The taxonomic relations are *isa* and *ako* (i.e. a kind of). Other types of relations (e.g. *has*, causal relations, compositional relations) are defined by object properties. Some examples of object properties are: *PossibleInfluencedBy*, *hasMeasureUnit*, *CausedBy*, *hasConcentration*, *hasEffect*, *hasGeneral-AQI*, *hasSpecific-AQI*, *hasSource*, *InfluencedBy*, *MayProduce*, *DefinedBy*. For example, *CausedBy*, *hasEffect* and *InfluencedBy* define cause-effect relations, while *PossibleInfluencedBy* is a possible causal relation. Examples of data properties are: *MeasurementUnit* (string), *AQI-Value* (byte), *AQI-ColourCode* (string), *ConcentrationStandardLimit* (real), *ConcentrationValue* (real).

Step 3. Knowledge Acquisition (for both problems, PA1 and PA2)

We have performed knowledge modelling for air pollution analysis - the case of acid rain occurrence in a sub-urban area and ozone prediction in an urban area. For each problem, we have taken from the conceptual model the corresponding sub-model, and the corresponding sub-ontology from AirPollution-Onto-1. Fig. 11 shows the acid rain sub-ontology graph, while Fig. 12 shows the ozone pollution sub-ontology graph, both depicted in OntoGraf.

Some examples of rules derived in this step for both problems, PA1, PA2, are given below. The certainty factors are set by using domain knowledge and human experts' advices. Also, a decision table (PA1-DT1) with selected examples is given in Table 6 for PA1 problem and another decision table (PA2-DT1) is given in Table 7.

Rule 12

IF AtmStability = True AND SO2(t) = Insignificant AND NOx(t) = Significant
AND Ozone(t) = Significant AND Ozone(t-1) = Significant AND Ozone(t-2) = Insignificant
THEN Ozone(t+1) = Significant CNF 91;

Rule A-2

IF PollutionSource=Traffic THEN PossibleAirPollutants = {PM_{10}, $PM_{2.5}$, CO, NO_x, VOC, less SO_2};

Rule A-3

IF PollutionSource=Industry THEN PossibleAirPollutants = {PM_{10}, $PM_{2.5}$, CO, NO_x, VOC, SO_2};

Rule A-4

IF PollutionSource=Domestic THEN PossibleAirPollutants={PM_{10}, $PM_{2.5}$, CO};

Rule T-7

IF NO_x = significant THEN PossibleEffects = {acid rain, smog};

Rule T-8

IF SO_2 = significant and O_3 = significant THEN PossibleEffect = increaseOzoneEffects;

Rule AQI-1

IF OzoneConc > 240 THEN AQI_O3 = 6; // according to Romanian air quality standard

Rule AR-3

IF NO_x = significant AND Precipitation=Significant AND SO_2 = Significant
THEN PollutionEffects = {AcidRain, WaterPollution, SoilPollution} CNF 98;

Rule A-26

IF CO_Conc > 10 THEN CO_S = high; // CO concentration [mg/m3]

Rule A-31

IF PM_{10}Conc < 50 THEN Effects = NoEffects; // 50 is the maximum allowed daily value [ug/m3]

Rule PA2-A1

IF clear_sky = true AND warm_temp = true AND light_wind = true
THEN stable_condition = true;

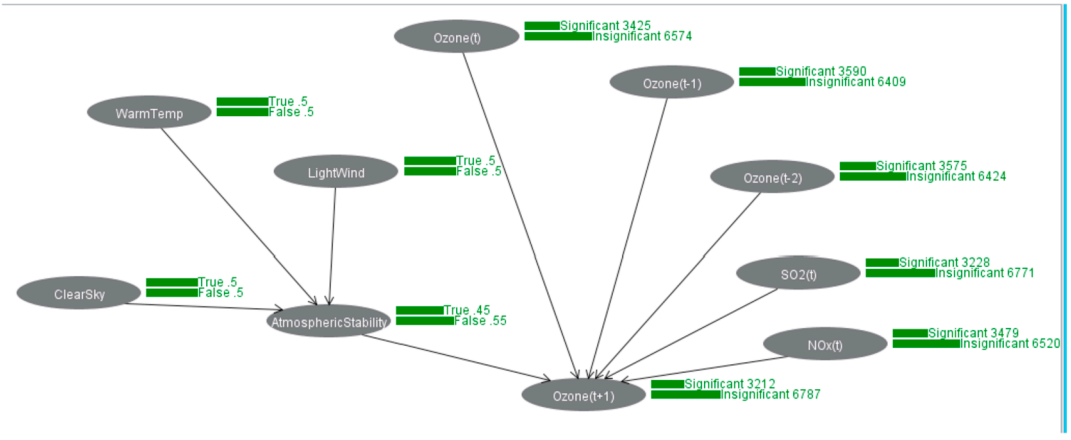

Fig. 14. The Bayesian network for Ozone Prediction scenario – PA2 (Weka 3.8.0).

Step 4. Knowledge discovery - via data mining (for PA1 and PA2)

During this step of the framework it is performed knowledge discovery starting with data cleaning and pre-processing and continuing with inductive data mining on cleaned and pre-processed data sets.

The datasets used for this case study were taken from the Romanian Air Quality Monitoring Network (RNMCA) site (www.calitateaer.ro), which are public available. They include hourly measurements of air pollutants concentration and meteorological parameters at some urban monitoring stations. The data were cleaned by using the graphical visualization tools of Weka. Data pre-processing included missing values imputation or removal, noise removal with filters and data normalization. Inductive rule learning was used to identify the air pollutants that were most correlated to ozone. The inductive DM algorithms that were experimented are M5P, REPTree, RandomForest (decision tree based algorithms) and M5Rules, Decision Tables (rule based algorithms). M5Rules classifier provided the best results (compared to Decision Tables, REPTree and RandomForest), showing that for ozone the most correlated air pollutants are NO_x and SO_2 with a slighter influence of CO and PM_{10}, the correlation coefficient being 0.9121 with RMSE = 6.6928. In case of PA2, the proper number of past hours measurement of ozone was identified by using data mining (M5P classifier gave the best correlation coefficient, 0.8861 for k = 2, past hours, and minimum root mean square error, RMSE = 7.8987).

Step 5. Bayesian network development (PA1, PA2)

One scenario was identified for each problem (PA1 and PA2): the acid rain scenario and the next hour ozone prediction scenario.

The corresponding Bayesian networks were developed in Weka and are shown in Figs. 13 and 14. The proper number of past hours measurement of ozone was identified by using the results of step 4 given by M5P classifier from Weka data mining software package.

As CO_2 is usually correlated with human activities (domestic sources) due to combustion of fossil fuels and we have used data from traffic or industrial sites situated in Ploiesti, in locations where CO_2 concentration has lower values, we have eliminated this air pollutant from the acid rain scenario.

The conditional probability tables of the Bayesian networks nodes were set from decision tables provided in step 3, according to problem domain knowledge.

Step 6. Knowledge acquisition from Bayesian networks

From the Bayesian network developed at the previous step we have derived the probabilistic rules. An examples of such rules is given below. The probabilities are derived from Bayesian networks.

Rule 7
IF ClearSky = True AND LightWind = True AND WarmTemp = True THEN AtmStability = True *P* (0.45);

Step 7. Knowledge integration and validation

The knowledge derived in the previous steps are integrated into the final knowledge base KB_{PA1} according to the algorithm given in section 3.

4.3. Case study 3 – environmental domain: soil

The last case study consider the soil pollution problem.
Problem (PS1): Soil pollution;
Problem description:
Soil pollution in a certain geographical area (e.g. urban, suburban, rural) can be caused by: industry, waste, air pollution, acid rain, water pollution, agriculture and agrochemicals, transportation, and other sources (e.g. specific to the geographical area). For example, industry can pollute the soil with heavy metals, methane, PAH, petroleum hydrocarbons etc. Also, agriculture through the use of different agrochemicals such as pesticides, herbicides and fertilizers can strongly increase the soil pollution level.

Step 1. Conceptual model design

The conceptual model for PS1 (CM_{PS1}) is depicted in Fig. 15.

Step 2. Ontology development

A vocabulary of terms and relations between them was defined and

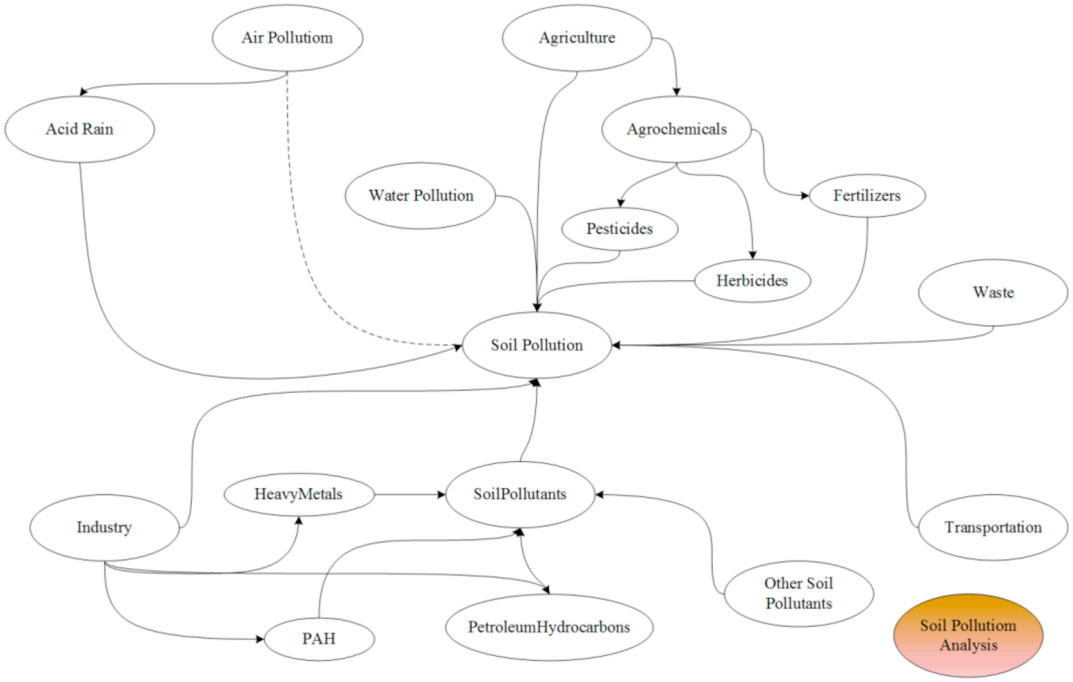

Fig. 15. The conceptual model for soil pollution analysis problem (CM_{PS1}).

Fig. 16. Sub-classes from Soil Pollution Ontology prototype in OWL Viz (Protégé 4.3).

Table 8
PS1-DT1 decision table (selection).
Step 4. Knowledge discovery - via data mining

GroundOzone	AcidRain	HeavyMetals	Waste	SoilPollution
Yes	No	No	No	Yes CNF 75
Yes	Yes	No	No	Yes CNF 84
Yes	Yes	Yes	No	Yes CNF 95
Yes	Yes	Yes	Yes	Yes CNF 100
No	Yes	No	No	Yes CNF 65
No	No	Yes	No	Yes CNF 90
No	No	No	Yes	Yes CNF 70
No	No	No	No	No

a prototype Soil Pollution ontology was implemented in Protégé 4.3. Fig. 16 presents two screenshots with concepts sub-classes in OWL Viz for soil pollutants, and soil remediation.

Step 3. Knowledge Acquisition

Some rules and decision tables were derived from literature and human experts. The decision table, PS1-DT1 (the version with binary values - Yes/No) is given in Table 8 and some examples of rules (with nominal values) are given below.

Rule S-1
IF AcidRain = True THEN SoilPollution = True CNF 84;
Rule S-2
IF GroundOzone = Moderate THEN SoilPollution = True CNF 75;
Rule 14
IF Industry = {WoodInd, Mining, Mettalurgy, ChemicalInd, PlasticsInd, ElectronicInd}
THEN SoilPollutants = HeavyMetals;

During this framework step of knowledge discovery it is performed data cleaning and pre-processing followed by inductive data mining on cleaned and pre-processed data sets.

Due to lack of enough data sets for this case study, we have performed inductive learning only from decision tables, which did not required data cleaning and pre-processing.

Step 5. Bayesian network development

PS1 Problem's Scenario.

One scenario was identified for the PS1 problem: methane and heavy metals related soil pollution. Fig. 17 presents the *soil pollution due to methane and heavy metals scenario sub-ontology* graph (in OntoGraf).

We have implemented a Bayesian network for soil pollution analysis due to ground ozone, acid rain, waste and heavy metals, shown in Fig. 18. For a more complex analysis, we can integrate in the Soil Pollution Bayesian network, the Bayesian network of AcidRain, given in

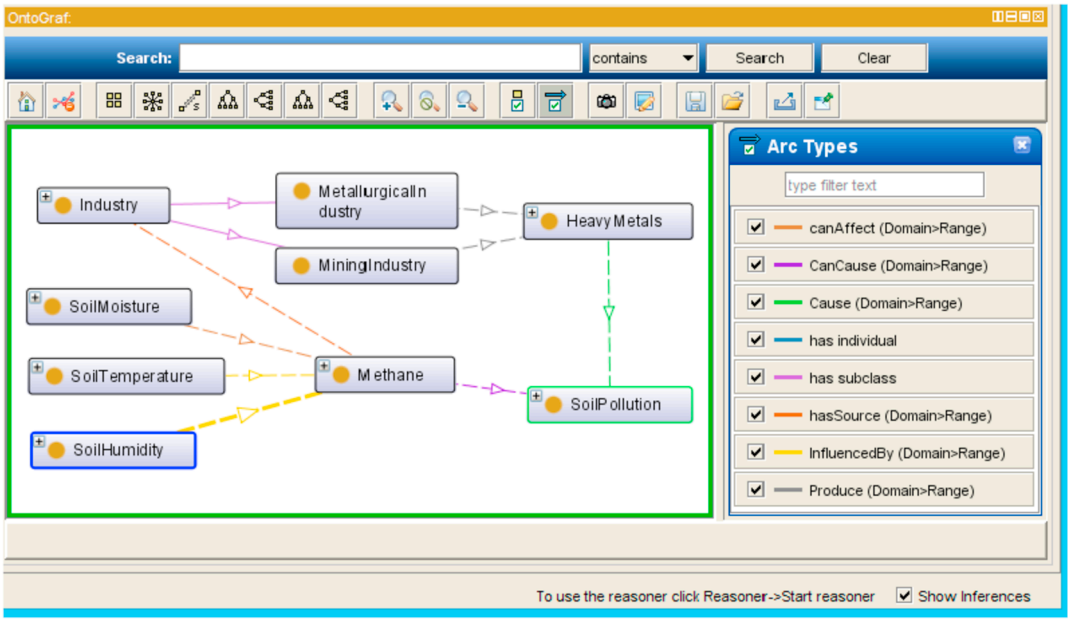

Fig. 17. Soil Pollution Onto-1 soil pollution due to methane and heavy metals scenario sub-ontology graph (in OntoGraf).

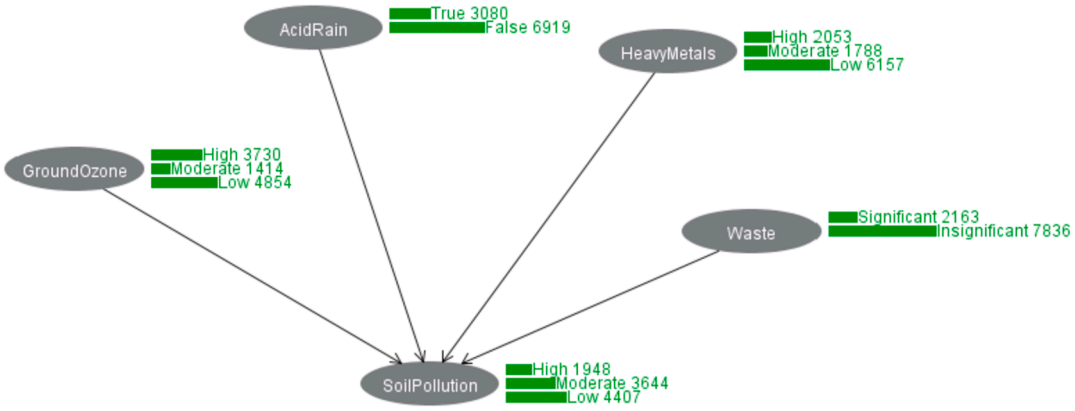

Fig. 18. Bayesian network for Soil Pollution scenario (Weka 3.8.0).

Fig. 13, and that of ozone prediction, given in Fig. 14, providing a more complex Soil Pollution Bayesian network, with a modular structure.

Step 6. Knowledge acquisition from Bayesian networks

From the Bayesian network developed at the previous step we have derived the probabilistic rules. Examples of such rules are given below. The probabilities are derived from Bayesian networks.

Rule S-3

IF GroundOzone = High $P(0.37)$ AND AcidRain = True $P(0.30)$ AND HeavyMetals = High $P(0.20)$ AND Waste = Significant $P(0.21)$ THEN SoilPollution = High $P(0.19)$; // derived from BN

Step 7. Knowledge integration and validation

The knowledge derived in the previous steps are integrated into the final knowledge base KB_{PS1} according to the algorithm given in section 3.

5. Conclusion

Environmental knowledge modelling can benefit from the integration of data and knowledge driven approaches, as we proposed in the framework introduced in this paper. Our solution combines an ontological approach (knowledge driven) with two analysis approaches (data mining - data driven, and Bayesian networks - data and knowledge driven) for the generation of a knowledge base for an IEDSS. Three case studies for different environmental problems (water resource management, water pollution analysis, air pollution analysis, ozone prediction and soil pollution analysis) were described.

The main advantage of the environmental knowledge modelling framework is that of dealing with complex environmental problems in an incremental and modular way by smaller scenarios analysis (see e.g. air and soil pollution analysis for acid rain and ozone/ground level ozone) supervised by the problem domain ontology. Each scenario is built starting from the conceptual model, using the ontology and knowledge derived from literature/human experts or via data mining (of quantitative and qualitative data), and implemented in Bayesian networks, which gives the probabilistic rules for decision making. Knowledge uncertainty is quantified by certainty factors and probabilities as well as in the fuzzy terms included in rules.

As a future work we shall investigate the possibility of generating

semi-automatically the rules derived from Bayesian networks and we shall make a unification of the two uncertainty models, probabilistic and uncertainty factors, in order to have uniformity in rules description.

References

Aquillera, P.A., et al., 2011. Bayesian networks in environmental modelling. Environ. Model. Software 26, 1376–1388.

Armstrong, L.J., Diepeveen, D., Maddern, R., 2007. The application of data mining techniques to characterize agricultural soil profiles. In: The 6th Australasian Data Mining Conference (AusDM 2007), Conferences in Research and Practice in Information Technology (CRPIT), vol. 70. pp. 81–96.

Babovic, V., Drécourt, J.-P., Keijzer, M., Hansen, P.F., 2002. A data mining approach to modelling of water supply assets. Urban Water 4 (4), 401–414.

Barton, D.N., Kuikka, S., Varis, O., Uusitalo, L., Henriksen, J., Borsuk, M., de la Hera, A., Farmani, R., Johnson, S., Linnell, J.D.C., 2012. Bayesian networks in environmental resource management. Integrated Environ. Assess. Manag. 8 (3), 418–429.

Beck, H., Morgan, K., Jung, Y., Grunwald, S., Kwon, H.-Y., Wu, J., 2010. Ontology-based simulation in agricultural systems modelling. Agric. Syst. 103, 463–477.

Breiman, L., Friedman, J.,H., Olshen, R.A., Stone, P.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Global Biogeochem. Cycles 23https://doi.org/10.1029/2009GB003506. GB4033.

Carmona, G., Varela-Ortega, C., Bromley, J., 2011. The use of participatory object-oriented Bayesian networks and agro-economic models for groundwater management in Spain. Water Resour. Manag. 25, 1509–1524. https://doi.org/10.1007/s11269-010-9757-y.

Ceccaroni, L., Cortés, U., Sànchez-Marrè, M., 2004. OntoWEDSS: augmenting environmental decision-support systems with ontologies. Environ. Modelling Sci. 19, 785–797.

Chau, K.W., 2007. An ontology-based knowledge management system for flow and water quality modeling. Adv. Eng. Software 38, 172–181.

Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environ. Model. Software 37, 134–145.

Chen, S.H., et al., 2008. Artificial intelligence techniques: an introduction to their use for modelling environmental systems. Math. Comput. Simulat. 78, 379–400.

Cioaca, E., Linnebank, F.E., Bredeweg, B., Salles, P., 2009. A qualitative reasoning model of algal bloom in the Danube Delta Biosphere Reserve (DDBR). Ecol. Inf. 4 (5–6), 282–298.

Cios, K.J., et al., 2007. Data Mining - a Knowledge Discovery Approach. Springer.

Clark, P., Niblett, R., 1989. The CN2 inductive algorithm. Mach. Learn. 3, 261–284.

Corani, G., Scanagatta, M., 2016. Air pollution prediction via multi-label classification. Environ. Model. Software 80, 259–264.

Cossentino, M., Raimondi, F.M., Vitale, M.C., 2001. Bayesian models of the PM_{10} atmospheric urban pollution. Trans. Ecol. Environ. 47, 143–152. https://doi.org/10.2495/AIR010141.

Czarnecki, A., Orłowski, C., 2009. Hybrid Approach to Ontology Specification and Development. pp. 47–57. https://doi.org/10.13140/2.1.3914.5762.

Czechowski, P., Badyda, A., Majewski, G., 2013. Data mining system for air quality monitoring networks. Arch. Environ. Protect. 39 (4), 123–144. https://doi.org/10.2478/aep-2013-0041.

Deb, C.K., Marwaha, S., Malhotra, P.K., Wahi, S.D., Pandey, R.N., 2015. Strengthening soil taxonomy ontology software for description and classification of USDA soil taxonomy up to soil series. In: 2nd Int. Conf. On Computing for Sustainable Global Development (INDIACom). IEEE.

Du, F., Zhu, A.-X., Band, L., Liu, J., 2014. Soil Property Variation Mapping through Data Mining of Soil Category Maps, Hydrological Processes. John Wiley & Sons, Ltdhttps://doi.org/10.1002/hyp.10383.

Dutta, R., et al., 2014. Development of an intelligent environmental knowledge system for sustainable agricultural decision support. Environ. Model. Software 52, 264–272.

Ekasingh, B., Ngamsomsuke, K., Letcher, R.A., Spate, J., 2005. A data mining approach to simulating farmers' crop choices for integrated water resource management. J. Environ. Manag. 77, 315–325.

Fenz, S., Min Tjoa, A., Hudec, M., 2009. Ontology-based Generation of Bayesian Networks, International Conference on Complex. Intelligent and Software Intensive Systems, pp. 712–717.

Friedman, J.H., 1977. A recursive partitioning decision rule for non-parametric classification. IEEE Trans. Comput. 404–408.

Garrido, J., Requena, I., 2011. Proposal of ontology for environmental impact assessment: an application with knowledge mobilization. Expert Syst. Appl. 38, 2462–2472.

Gibert, K., Sànchez-Marrè, M., Codina, V., 2010a. Choosing the right data mining technique: classification of methods and intelligent recommenders. Proceedings of iEMS 2010 (1), 2448–2453.

Gibert, K., Rodríguez-Silva, G., Rodríguez-Roda, I., 2010b. Knowledge discovery with clustering based on rules by states: a water treatment application. Environ. Model. Software 25 (6), 712–723.

Gibert, K., Sànchez-Marrè, M., Sevilla, B., 2012. Tools for environmental data mining and intelligent decision support. In: Proceedings of IEMS 2012.

Gibert, K., Sànchez-Marrè, M., Izquierdo, J., 2016. A survey on pre-processing techniques: relevant issues in the context of environmental data mining. AI Communications 29 (6), 627–663.

Heller, U., Struss, P., 2001. Transformation of qualitative dynamic models - application in hydroecology. In: Hotz, L., Struss, P., Guckenbiehl, T. (Eds.), Intelligent Diagnosis in Industrial Applications. Shaker Verlag, Aachen, Germany, pp. 95–106.

Huang, J.J., McBean, E.A., 2009. Data mining to identify contaminant event locations in water distribution systems. J. Water Resour. Plann. Manag. 135 (6), 466–474.

Karimipour, F., Delavar, M.R., Kinaie, M., 2005. Water quality management using GIS data mining. J. Environ. Inform. 5 (2), 61–72.

Kochilakis, G., et al., 2016. A web based DSS for the management of floods and wildfires (FLIRE) in urban and periurban areas. Environ. Model. Software 86, 111–115.

Kolli, K., Seshadri, R., 2013. Ground water quality assessment using data mining techniques. Int. J. Comput. Appl. 76 (15), 39–45.

Kusiak, A., 2002. Data mining and decision making. In: In: Dasarathy, B.V. (Ed.), Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV, vol. 4730. SPIE, Orlando, FL, pp. 155–165.

Kwon, H.-Y., Grunwald, S., Beck, H.W., Jung, Y., Daroub, S.H., Lang, T.A., Morgan, K.T., 2010. Ontology-based simulation of water flow in organic soils applied to Florida sugarcane. Agric. Water Manag. 97, 112–122.

Laniak, G.F., et al., 2013. Integrated environmental modelling: a vision and roadmap for the future. Environ. Model. Software 39, 3–23.

Li, S.-T., Shue, L.-Y., 2004. Data mining to aid policy making in air pollution management. Expert Syst. Appl. 27 (3), 331–340.

Li, X., Xie, Y., Li, L., Yang, X., Wang, N., Wang, J., 2015. Using robust Bayesian network to estimate the residuals of fluoroquinolone antibiotic in soil. Environ. Sci. Pollut. Res. 22, 17540–17549. https://doi.org/10.1007/s11356-015-4751-9.

Liu, K.F.-R., Lu, C.-F., Chen, C.-W., Shen, Y.-S., 2012. Applying Bayesian belief networks to health risk assessment. Stoch. Environ. Res. Risk Assess. 26, 451–465.

Ma, Y., Richards, M., Ghanem, M., Guo, Y., Hassard, J., 2008. Air pollution monitoring and mining based on sensor grid in London. Sensors 8, 3601–3623. https://doi.org/10.3390/s8063601.

McDonald, K.S., et al., 2016. An ecological risk assessment for managing and predicting trophic shifts in estuarine ecosystems using a Bayesian network. Environ. Model. Software 85, 202–216.

McIntosh, B.S., et al., 2011. Environmental decision support systems (EDSS) development – challenges and best practices. Environ. Model. Software 26, 1389–1402.

Metral, C., Falquet, G., Karatzas, K., 2008. Ontologies for the integration of air quality models and 3D city models. In: 2nd Workshop COST Action C21-towntology, Ontologies for Urban Development: Conceptual Models for Practitioners, pp. 18–33.

Oprea, M., 2005. A case study of knowledge modelling in an air pollution control decision support system. AI Communications 18 (4), 293–303.

Oxley, T., et al., 2004. Integrated modelling and decision support tools: a Mediterranean example. Environ. Model. Software 19, 999–1010.

Pérez-Miñana, E., Krause, P.J., Thornton, J., 2012. Bayesian networks for the management of greenhouse gas emissions in the British agricultural sector. Environ. Model. Software 35, 132–148. https://doi.org/10.1016/j.envsoft.2012.02.016.

Phan, T.D., et al., 2016. Applications of Bayesian belief networks in water resource management: a systematic review. Environ. Model. Software 85, 98–111.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1 (1), 81–106.

Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343–348.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

Riga, M., Tzima, F.A., Karatzas, K., Mitkas, P.A., 2009. Development and evaluation of data mining models for air quality prediction in Athens, Greece. In: Athanasiadis, I.N. (Ed.), Information Technologies in Environmental Engineering, Environmental Science and Engineering. Springer-Verlag, pp. 331–344. https://doi.org/10.1007/978-3-540-88351-7_25.

Robertson, D.E., Wang, Q.J., Malano, H., Etchells, T., 2009. A Bayesian network approach to knowledge integration and representation of farm irrigation: 2. Model Validation, Water Resource Res. 45 (2). https://doi.org/10.1029/2006WR005420.

Sànchez Marrè, M., et al., 2008a. Towards a framework for the development of intelligent environmental DSSs. In: Proceedings of iEMS 2008, 1, pp. 398–406.

Sànchez Marrè, M., Gibert, K., Sojda, R.S., Steyer, J.P., Struss, P., Rodriguez-Roda, I., Comas, J., Brilhante, V., Roehl, E.A., 2008b. Intelligent environmental decision support systems. In: In: Jakeman, A., Rizzoli, A., Voinov, A., Chen, S. (Eds.), State of the Art and Futures in Environmental Modelling and Software, vol. 3. Elsevier Science, Amsterdam, The Netherlands, pp. 119–144 chap. 8.

Sánchez-Alonso, S., Sicilia, M.-A., 2009. Using an AGROVOC-based Ontology for the Description of Learning Resources on Organic Agriculture, Chapter in Book: Metadata and Sema ntics. Springer, pp. 481–492.

Siwek, K., Ossowski, S., 2016. Data mining techniques for prediction of air pollution. Int. J. Appl. Math. Comput. 26 (2), 467–478.

Sokolova, M.V., Fernández-Caballero, 2007. A multi-agent architecture for environmental impact assessment: information fusion, data mining and decision making. In: Proceedings of ICEIS 2007, vol. 2. pp. 219–224.

Stanley Young, S., Xia, J.Q., 2013. Assessing geographic heterogeneity and variable importance in an air pollution data set. Stat. Anal. Data Min. 6, 375–386.

Struss, P., 2008. Artificial intelligence based modeling for environmental applications and decision support. In: Proceedings of International Congress on Environmental Modelling and Software Society. Catalonia, Spain, Barcelona.

Struss, P., Bendati, M., Lersch, E., Roque, W., Salles, P., 2003. Design of a model-based decision support system for water treatment. In: Proceedings of the 18th IJCAI 03-Environmental Decision Support Systems Workshoppp. 50–59 Acapulco, Mexico).

Tang, C., Yi, Y., Yang, Z., Sun, J., 2016. Risk analysis of emergent water pollution accidents based on a Bayesian network. J. Environ. Manag. 165, 199–205.

Teixeira, S., Guimarães, A.M., Proença, C.A., da Rocha, J.C.F., Caires, E.F., 2014. Data mining algorithms for prediction of soil organic matter and Clay based on vis-NIR spectroscopy. Int. J. Agric. For. 4 (4), 310–316.

Urbani, D., Delhom, M., 2005. Water management policy selection using a decision support system based on multi-agent systems. LNCS 3673, 466–469.
Wang, Y., Witten, I.H., 1997. Induction of model trees for predicting continuous classes. In: Proceedings of the 9th European Conference on Machine Learning.
Wang, Q.J., Robertson, D.E., Haines, C.L., 2009. A Bayesian network approach to knowledge integration and representation of farm irrigation: 1. Model development, Water Resource Research 45 (2). https://doi.org/10.1029/2006WR005419.
Wen, X.-P., Yang, X.-F., 2011. Monitoring water quality using remote sensing data mining. In: Funatsu, Kimito (Ed.), Chapter in Book: Knowledge Applications in Data Mining. InTech ISBN: 978-953-307-154-1.
Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, third ed. Elsevier.
Wotawa, F., 2011. On the use of abduction as an alternative to decision trees in environmental decision support systems. Int. J. Agric. Environ. Inf. Syst. 2 (1), 63–82.
Wotawa, F., Rodriguez-Roda, I., Comas, J., 2010. Environmental decision support systems based on models and model-based reasoning. Environ. Eng. Management J. 9 (2), 189–195.
Xiaomin, Z., et al., 2016. An ontology-based knowledge modelling approach for river water quality monitoring and assessment. Proc. Comp. Sci. 96, 335–344.